博　士　論　文
Doctoral Dissertation

Visual Understanding of Human Hands in Interactions

(インタラクションにおける手の視覚的理解)

大川武彦 (Takehiko Ohkawa)

Advisor: Professor Yoichi Sato

Department of Information and Communication Engineering

Graduate School of Information Science and Technology

The University of Tokyo

June 2025

In every sense, it's a story of hands — hands that were studied with care, hands that worked with devotion, and hands that reached out with kindness.

Abstract

Human hand interactions are central to daily activities and communication, providing informative signals for human action, expression, and intent. Visual perception and modeling of hands from images and videos are therefore crucial for various applications, including understanding human interactions in the wild, virtual human modeling in 3D, human augmentation through assistive vision systems, and highly dexterous robotic manipulation.

While computer vision has evolved to estimate hand states ranging from coarse detection to nuanced 3D pose and shape estimation, existing methods fall short in three key challenges. First, they struggle to handle complicated and fine-grained contact scenarios, such as grasping objects (*i.e.*, hand-object contact) or touching one's own body (*i.e.*, self-contact), where occlusions and deformations introduce substantial ambiguity. Second, most models generalize poorly to dynamic and real-world environments due to the domain gap between studio-collected training datasets and in-the-wild testing conditions. Third, beyond geometric estimation, current approaches often lack the ability to link low-level hand states to high-level semantic comprehension, *e.g.*, with action labels or language.

This dissertation addresses these limitations by pursuing the goal of **precise tracking and interpretation of fine-grained hand interactions from real-world visual data**. To achieve this, the dissertation systematically proposes three key pillars:

- **Data foundation**: Building diverse and high-quality data infrastructure to enable learning fine-grained hand interactions featuring challenging contact scenarios.
- **Robust modeling for fine details**: Achieving robust and reliable estimation for fine-grained hand interactions by generalizing and adapting machine learning models, making them resilient to occlusion, noise, and domain shift in in-the-wild scenarios.
- **Connecting geometry and semantics**: Bridging captured geometric information with semantics to comprehend actions and intentions based on tracked hand states.

The *first pillar* of this research focuses on building a diverse and high-quality data foundation. This involves investigating and capturing hand interaction datasets that include complex contacts, such as object interaction and self-contact, using multi-camera systems. This approach enables high-precision 3D pose and shape annotations for intricate hand contact, while offering valuable assets to the community.

The *second pillar* is dedicated to robust modeling to capture fine details. Leveraging the constructed datasets, we develop advanced machine learning methods for generalizable and adaptable estimation in the wild. This includes comprehensive analysis of 3D hand pose estimation tasks during object contact, building model priors from diverse image or pose data for downstream tasks in pose estimation, and proposing adaptation methods to further bridge performance gaps across different recording environments and camera settings.

The *third pillar* centers on connecting the captured low-level geometric information with high-level semantic understanding. This involves utilizing the predictions for hand geometry in 2D or 3D (e.g., detection, segmentation, and pose) to comprehend the semantics of the interaction. We explore natural language descriptions as semantic signals and propose generating dense video captions from the hand-object tracklets.

Collectively, these three pillars present a consistent and integrated framework to advance the visual understanding of hands in interactions. By combining advanced techniques in data foundation, robust modeling, and semantic understanding, this dissertation contributes to foundational technologies and intelligent systems for human-centric interactions, with broader applications and implications in computer vision.

Acknowledgments

First and foremost, I am deeply indebted to my advisor, Prof. Yoichi Sato, for his invaluable guidance and unwavering support. His enthusiasm, perseverance, and thoughtful mentorship have shaped both my academic and personal growth, inspiring me to cultivate a pioneering, collaborative, and open-minded spirit. I am also deeply thankful to Ryosuke Furuta for his continuous advice and support in our daily discussions, paper reviews, and laboratory management.

My sincerest gratitude goes to the members of my dissertation committee, namely Prof. Toshihiko Yamasaki, Prof. Shin'ichi Satoh, Prof. Yusuke Sugano, and Prof. Yusuke Matsui. Their thorough evaluation of my work and insightful suggestions strengthened this dissertation. I also thank Prof. Kiyoharu Aizawa for his constructive and practical advice at the beginning of my PhD studies.

I would like to extend my sincere thanks to my collaborators and mentors from research institutions around the world for their generous support and fruitful collaboration. At Carnegie Mellon University (CMU), I thank Prof. Kris Kitani, Yu-Jhe Li, Qichen Fu, and Shun Iwase for their valuable input, stimulating discussions, and feedback during my visit. At ETH Zurich, I am especially thankful to Prof. Marc Pollefeys and Taein Kwon, whose guidance during my stay greatly enriched my technical perspective and research direction. I appreciate the industrial collaboration and support at Meta, specifically from Kun He, Takaaki Shiratori, Shunsuke Saito, Jason Saragih, Jihyun Lee, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. My long-term collaboration with OMRON SINIC X, together with Takuma Yagi, Atsushi Hashimoto, Yoshitaka Ushiku, and Taichi Nishimura, was invaluable in providing broader insights into building foundational datasets and practical vision systems.

My sincere thanks go to the academic consortium for hands research, including Prof. Linlin Yang, Zicong Fan, Prof. Angela Yao, Rongyu Chen, Prof. Lixin

Yang, Prof. Qi Ye, Prof. Hyung Jin Chang, Prof. Otmar Hilliges, Aditya Prakash, and Prof. Saurabh Gupta, among others, for their efforts in hosting international workshops together at ICCV 2023, 2025 and ECCV 2024. This fosters next-generation research and advances the collective knowledge in the field. In particular, I thank Prof. Angela Yao and Prof. Saurabh Gupta for inviting me to give seminar talks at the National University of Singapore (NUS) in 2023 and the University of Illinois Urbana-Champaign (UIUC) in 2025, respectively.

I am grateful for the financial support that enabled my research activities, including international research stays. This includes funding from JST ACT-X Project (2020–2023), JSPS Research Fellowship (DC1) (2022–2024), and fellowships from Google PhD Fellowship in the Machine Perception track (2024), Microsoft Research Asia (2023), and ETH Zurich Leading House Asia (2023). I sincerely thank Prof. Imari Sato, Prof. Ken-ichi Kawarabayashi, Masataka Goto, and Prof. Nakamasa Inoue for their persistent guidance on the JST ACT-X project over nearly four years. I also appreciate the mentorship and periodic discussions with my industrial research advisors through fellowship programs, namely Jinglu Wang and Mawo Kamakura at Microsoft Research Asia and Yasuhisa Fujii at Google DeepMind.

I am equally thankful to all the members of the Sato/Sugano Lab, with whom I have had the pleasure of working over the years. I am grateful to Nie Lin, Ruicong Liu, Yilin Wen, Tatsuro Banno, and Naru Suzuki for inspiring conversations that led us into human hand modeling in 3D. My thanks also go to Yifei Huang, Takumi Nishiyasu, Wataru Kawabe, Mingfang Zhang, Masatoshi Tateno, Liangyang Ouyang, Chiyun Li, Yuki Maeda, and Zhifan Zhu for encouraging daily discussions and shared passion for research and life.

I gratefully acknowledge the GPU hardware donation supported by renowned MLB player Yu Darvish. This enabled me to broaden my global perspective and pursue research at the frontier, as he has pursued in his own baseball career.

I would also like to express my heartfelt gratitude to my family for their encouragement, patience, and belief in me. Their support has been the cornerstone of my academic journey. Last but not least, I thank Ayano, my beloved partner, for all the *live, laugh, love* you bring into my world.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Ubiquitous Role of Hands

> "*The hand is the cutting edge of the mind.* Civilisation is not a collection of finished artefacts, it is the elaboration of processes. In the end, the march of man is the refinement of the hand in action."
> — Jacob Bronowski [34]

Human hands play a unique and indispensable role in shaping our reality and defining our interactions, as encapsulated by the profound observation from mathematician Jacob Bronowski [34]. Far more than mere manipulators, hands serve as a primary interface between an individual and their surroundings, facilitating engagement with the self, the physical world, and the rapidly expanding virtual realm. We intuitively grasp this centrality of hands, recognizing their influence from our earliest experiences through every stage of life, as shown in Fig. 1.1.

From our very first moments, our hands are our primary explorers. As a child, hands are fundamental to our cognitive and physical growth. They inform how we first interact with the world around us, reaching out to grasp objects, learning about textures, shapes, and gravity. These early interactions, mediated solely by our hands, facilitate the understanding of cause and effect, the development of motor skills, and the building of our sensory perception. Long before words, our hands serve as our first language, expressing needs and curiosity.

As we grow, our hands become powerful instruments of connection and communication. When we interact with friends, family, or even strangers, our hands

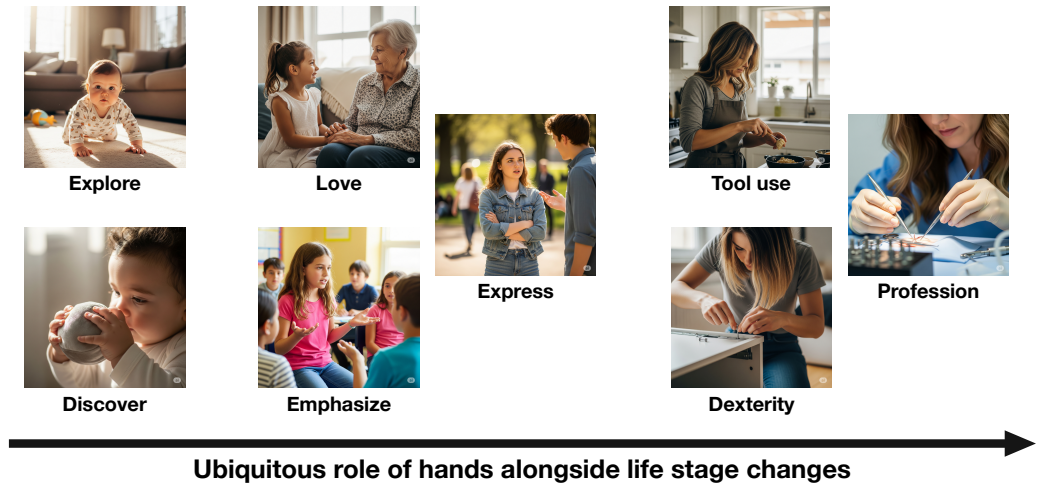


Figure 1.1: **Ubiquitous role of human hands**. The figure illustrates a conceptual framework depicting the ubiquitous role of human hands across various life stages, from infancy to adulthood. Each panel showcases a distinct hand-centric activity, including exploration, affection, expression, tool utilization, dexterity, and professional activity, thereby demonstrating the continuous evolution of hand functions as individuals progress through different developmental phases. All images are generated by an AI assistant [8].

can convey emphasis, emotion, and intent. Beyond spoken words, gestures emphasize a point, express joy or frustration, and build rapport through a handshake or a comforting touch. For communities, such as those using sign language, our hands facilitate complex linguistic and emotional exchanges, showcasing their capacity for nuanced communication.

As we mature and engage in daily life or professional endeavors, the dexterity and adaptability of human hands become paramount. Whether preparing ingredients in cooking, assembling components in manufacturing, or executing microsurgery or biological experiments, our hands manipulate our environment with precise grasps or tool use, allowing us to accomplish these tasks. This capability for dexterous manipulation enables us to build, create, and master complex procedures, while realizing abstract thought into tangible reality.

In essence, hands are central to what it means to be human. They are not merely tools, but extensions of our minds, providing rich, informative signals that unveil not only what actions are being performed but also the underlying dynam-

ics, expressions, and immediate intents of an individual. This profound centrality makes the perception and modeling of hands a grand challenge, yet a highly rewarding endeavor, in the pursuit of human-centric intelligent systems.

1.2 Applications

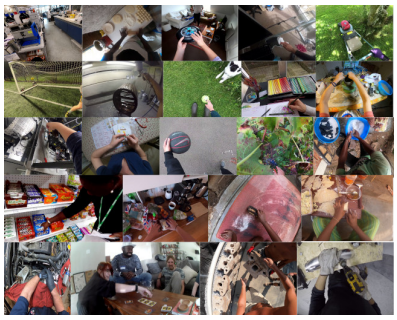
The ubiquitous role of hands, as established in the preceding section, directly translates into a critical need for advanced applications of visual perception and modeling of hand interactions in numerous domains. Particularly, the ability to accurately track and interpret hand interactions is essential for developing human-centric intelligent systems, including video understanding, augmented reality, virtual reality, assistive technologies, and robotics (see Fig. 1.2).

Video understanding and skill analysis: In the understanding of human-centric videos, tracking hands in videos provides crucial probes into *where* people interact and *what* they perform tasks [17, 109, 187], offering insight into actions, object interactions, and skills. This analysis particularly helps enhance activity understanding [77, 111] in complex tasks such as cooking [77, 257], assembly [252, 303], or biomedical procedures [247, 373].

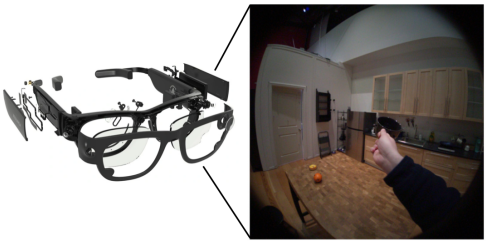
Beyond simple action recognition, capturing fine-grained kinematics of hands facilitates detailed skill analysis, providing objective metrics for assessing performance in fields like surgical training [29, 133, 401] and industrial quality control [35, 177]. For example, analyzing a surgeon's hand interactions can objectively quantify proficiency and identify areas for improvement, a capability increasingly relevant in medical education.

Augmented Reality (AR): A first-person (egocentric) perspective through AR glasses transforms how users interact with digital content overlaid on the real world. The field and research have been accelerating owing to recent commercial AR smart glasses including Magic Leap, XReal, Mixed Reality devices including Apple Vision Pro, Microsoft HoloLens, and open-source projects like Aria Glasses [85]. Combined with always-on visual sensing from the egocentric perspective, precise hand tracking in AR applications allows users to directly manipulate virtual objects, scroll through menus, or activate functions using natural gestures [36, 87, 212, 352].

Virtual Reality (VR): Similarly, in virtual reality, hand tracking systems elim-

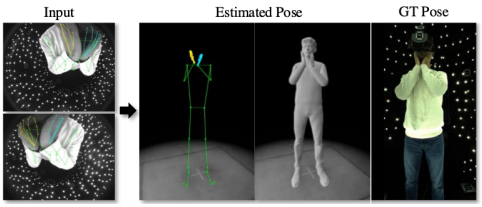


(a) **Video understanding, *e.g.*, Ego4D challenges [111].**



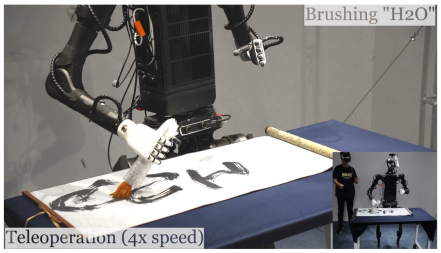
(b) **AR glasses and vision systems like Project Aria [85].**

(c) **VR game with hand tracking [119].**



(d) **Photorealistic avatar telepresence with VR [175].**

(e) **Assistive systems (*e.g.*, sign language understanding [387]).**



(f) **Robot imitation of human skill [132].**

Figure 1.2: **Applications for visual understanding of hands**.

inate the need for cumbersome physical controllers, and assist how users interact within purely digital environments [118, 119]. This progress is evident with the widespread commercial VR headsets like Meta Quest, Pico Neo, PlayStation VR, etc. Enabling users to employ their own hands as direct input devices opens applications for human-computer interaction, such as gaming [167], surface typing [287, 404], and virtual meetings [1]. This effectively enhances immersion, presence, and work efficiency, allowing for natural manipulation of virtual objects and intuitive navigation with hands. In VR telepresence, accurate human model-

ing including hands supports realistic avatar communication [175, 182, 215], conveying nuanced gestures and expressions that enrich the sense of connection and presence among remote participants.

Assistive technologies: The ability to precisely interpret hand interactions holds potential for assistive AI and accessibility applications. For deaf people, real-time sign language tracking and interpretation systems [6, 231, 342, 387] can translate gestures into spoken or written language, breaking down communication barriers. Beyond sign language, hand gesture recognition can provide intuitive control for smart home devices, wheelchairs, or digital interfaces for those with limited mobility [5]. This capability offers new avenues for interaction and independence that empower individuals with disabilities, thus promoting more inclusive technologies and extending to areas like rehabilitation [191, 394] for progress monitoring and feedback.

Robotics and human-robot collaboration: Finally, in robotics, understanding human hands is crucial for developing embodied, intelligent, and collaborative robotic systems. Precise hand tracking facilitates learning dexterous robot manipulation from human demonstration that would otherwise require extensive programming [75]. This capability is central to imitation learning [120, 132, 214, 276, 283, 406], where a robot acquires new skills by directly observing a human performing a task and replicating the intricate motions of human hands. Furthermore, in scenarios demanding human-robot collaboration [49, 142, 197], understanding hand gestures and intent allows robots to anticipate actions, avoid collisions, and work more safely and efficiently alongside human operators, impacting areas such as automated manufacturing and service robotics.

1.3 Challenges in Computer Vision

While the preceding section highlights the immense potential of visual hand understanding, realizing these applications confronts challenges in computer vision. While the field has made advances in estimating hand states, evolving from coarse detection [240, 305, 328] and segmentation [78, 256, 341, 400] to nuanced 3D pose and shape estimation [86, 88, 230, 421], current methods frequently fall short in real-world scenarios. This limitation stems from three key challenges: (i) the complexities of modeling fine-grained hand interactions, particularly in

contact-rich situations, (ii) the limited generalization ability of models trained in controlled environments to dynamic in-the-wild conditions, and (iii) the inherent difficulty in bridging the gap between low-level kinematics and high-level semantic meaning.

1.3.1 Vulnerability in contact

One of the persistent hurdles in visual hand understanding lies in accurately reconstructing hands during complicated, fine-grained contact scenarios. This problem is pronounced in two specific interaction types: *hand-object contact* and *self-contact*. In hand-object interactions, the object itself overlaps with the hand region, and object properties (*e.g.*, shape, size, transparency, deformability) also affect the hand's appearance. For self-contact, such as holding one's arm or touching the face, parts of the hand are inevitably occluded by other body parts, which introduces additional visual clutter and ambiguity. These difficulties can be summarized as follows:

- **Occlusion:** This is perhaps the foremost challenge. When a hand grasps an object or touches another part of the body, large portions of the hand, including crucial joints and fingertips, become self-occluded or object-occluded. This partial visibility leads to ambiguous observations, making precise 3D reconstruction exceptionally difficult for machine learning models.

- **Variability of contact:** The nature of contact is highly diverse—from a firm grasp to a light touch, a squeeze, or a sliding motion. Each type of contact affects the hand's deformation and appearance differently, requiring models to infer subtle changes in shape and pose that are not easily captured by standard skeletal models.

- **Difficulty of annotation:** Given these issues, annotating hand states (*e.g.*, mask or pose) for hand contact images is particularly challenging for machine-based annotation tools and even human annotators. This inherently limits the size and variety of available datasets, leading to dataset bias and narrow performance in different test domains.

1.3.2 Limited generalization to in-the-wild environments

Even with advances in hand modeling under controlled conditions, a fundamental challenge remains: machine learning models often fail to generalize to dynamic, real-world environments. This limitation is rooted in the *domain gap* between curated training datasets, typically captured in well-lit studios with fixed backgrounds, and the highly variable, cluttered testing conditions encountered in-the-wild, such as egocentric videos or AR/VR applications. Such generalization failures are especially problematic for practical deployment, where visual inputs are affected by uncontrolled lighting, motion blur, occlusions, and diverse backgrounds. The key obstacles to robust generalization are summarized as follows:

- **Domain discrepancy between laboratory and real-world settings:** Existing models are typically trained on clean, annotated datasets collected in laboratory environments with consistent lighting, limited background clutter, and synchronized multi-camera setups. However, these settings differ significantly from in-the-wild conditions, where hand appearances vary dramatically due to motion, camera viewpoint, clothing, lighting, and object diversity. As a result, models overfit to the specific biases of training datasets and perform poorly in unseen domains.

- **Supervision bottleneck in real-world data:** Capturing labeled data in-the-wild is expensive and labor-intensive, especially for 3D pose or contact annotations. Unlike studio setups with motion capture or multiview calibration, real-world environments lack such ground-truth supervision. This supervision bottleneck prevents supervised models from learning robust representations that can scale to uncontrolled inputs.

- **Need for pre-training and adaptation strategies:** To overcome these challenges, there is a growing need for scalable learning strategies that do not rely on dense supervision. This includes large-scale self-supervised pre-training on unlabeled human-centric videos, domain adaptation to mitigate the gap between source and target distributions, and robust prior modeling learned from diverse image or pose data. Such approaches help models to be resilient to appearance shifts, camera view changes, and environmental variability.

1.3.3 Gap in semantic comprehension

Beyond precise geometric tracking, a critical gap exists in connecting low-level geometric estimation of hand states (in 2D or 3D) to high-level semantic context. Current systems can often identify a hand's location and physical structure, but they struggle to interpret its semantics, such as performing specific actions, underlying affordance, expressing emotion, or conveying intent. Simply knowing the geometric structure (*e.g.*, the 3D pose of fingers) may not be sufficient for many downstream tasks that simultaneously require higher-level understanding. The main issues in connecting geometry to semantics are summarized as follows:

- **Lack of causal linkage:** Existing methods often lack explicit mechanisms to link low-level visual features and geometric hand information to high-level concepts such as action labels, procedures, or human intentions. Establishing causal modeling of these abstract concepts enables systems to predict future actions and provide adaptive intelligent assistance within a given context.

- **One-to-many semantic mapping:** A single observed geometric state (e.g., a specific hand pose or a short sequence of movements) can correspond to multiple different semantic meanings or be part of various higher-level activities. For instance, a hand forming a "grasping" configuration could be part of "drinking coffee," "picking up a tool," or "passing an object." The actual meaning is defined by the context, such as the sequence of geometric states, interaction with other entities (objects or body parts), broader environmental information, and the overarching human goal, not just the instantaneous configuration. This one-to-many mapping from geometry to semantic makes inference challenging.

- **Representations: Videos vs. Tracklets:** While *video-based* inputs offer rich spatiotemporal context from raw videos, they present challenges such as high-dimensional input size, increased computational complexity, and sensitivity to camera motion (*e.g.*, in dynamic egocentric videos). In contrast, *tracking-based* approaches, which process explicit hand tracklets (sequences of geometric states), can provide more compact and fine-grained representations. However, even with precise tracklets, a conclusive analysis

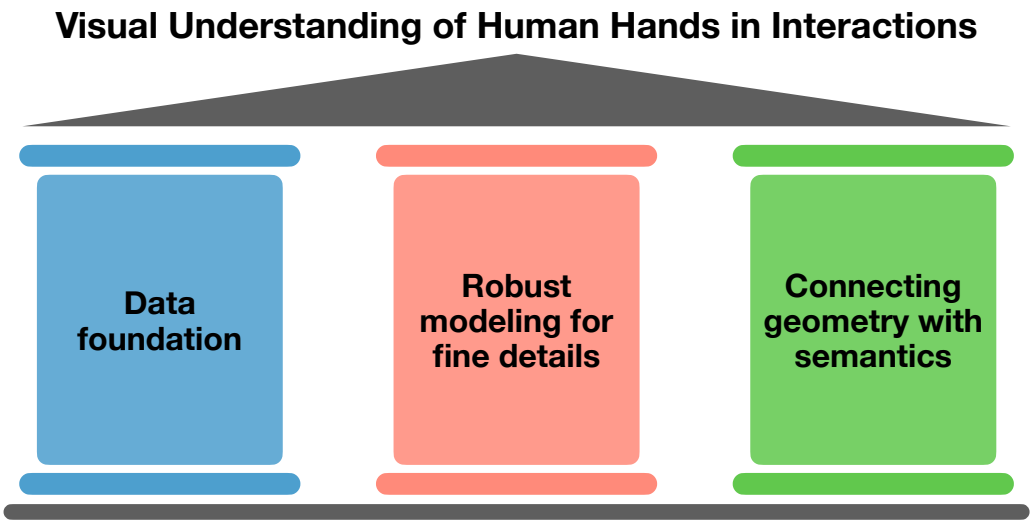


Figure 1.3: **Dissertation goal.** The figure provides the dissertation's strategic framework for achieving comprehensive visual understanding of hand interactions, systematically built upon three key pillars: *Data foundation*, *Robust modeling for fine details*, and *Connecting geometry with semantics.*

on the optimal representation for semantic comprehension, or their hybrid approach, remains an open research question.

1.4 Dissertation Goal

Building upon the identified challenges in the previous section, this dissertation sets forth a clear objective: **precise tracking and interpretation of fine-grained hand interactions from real-world visual data**. This ambitious endeavor aims to address the limitations and push the boundaries of human-centric computer vision, enabling a deeper and practical understanding of human behavior through the lens of hand interactions. By focusing on both the accurate geometric capture and the meaningful semantic interpretation of these interactions, this work lays the foundational groundwork for intelligent and intuitive human-AI systems.

1.5 Contributions

To achieve the goal of precise tracking and interpretation of fine-grained hand interactions, this dissertation is systematically built upon three key pillars, each addressing a fundamental challenge in the field. These three pillars refer to, namely, *Data foundation*, *Robust modeling for fine details*, and *Connecting geometry with semantics*, as illustrated in Fig. 1.3.

1.5.1 Pillar 1: Data foundation for hand interactions

Accurate modeling of complex hand interactions, particularly those involving contact, is inherently data-driven. Existing datasets often lack the diversity, scale, and precise annotations necessary to train robust models for these challenging scenarios, especially regarding occlusions and subtle deformations during contact. To overcome this, the first pillar of this research focuses on building a diverse and high-quality data infrastructure. This involves investigating and capturing novel hand interaction datasets that explicitly feature complex contacts, such as hand-object interactions and self-contact. We leverage multi-camera systems to enable high-precision 3D pose and shape annotations, capturing the intricate details of hand contact from multiple viewpoints. This meticulous data collection and annotation approach provides a rich, realistic, and large-scale foundation that is crucial for training robust models, thereby offering valuable assets to the research community and serving as solid benchmarks for future work.

1.5.2 Pillar 2: Robust modeling to capture fine details

Even with high-quality data, developing machine learning models that can robustly and reliably estimate fine-grained hand interactions in unconstrained "in-the-wild" scenarios remains a considerable hurdle. Challenges like varying lighting, diverse backgrounds, and dynamic camera movements introduce noise and ambiguity that can degrade model performance. Leveraging the constructed datasets, the second pillar of this dissertation focuses on developing advanced machine learning methods for generalizable and adaptable estimation in the wild. This includes a comprehensive analysis of 3D hand pose estimation, with a particular focus on hand-object contact scenarios. We also contribute by building self-

supervised model priors from diverse image or pose data, including large-scale pre-training and generative prior modeling. These priors can capture intrinsic hand properties and common interaction patterns from unlabeled data. Furthermore, we propose a novel adaptation method with self-training that bridges performance gaps across different recording environments and camera settings. This ensures that our models maintain accuracy and reliability even when deployed in novel and unconstrained conditions.

1.5.3 Pillar 3: Connecting geometry with semantics

The utility of precise hand tracking extends beyond mere geometric reconstruction; it lies in interpreting the meaning behind the movements. As discussed, a fundamental gap exists between low-level kinematics and high-level semantic understanding, limiting the ability of systems to comprehend human actions and intentions. The third pillar of this research centers on bridging this gap by connecting captured geometric information with semantic understanding. This involves effectively utilizing the predictions for geometry in 2D or 3D (*e.g.*, detection, segmentation, and pose) to infer the broader semantics of the interaction. We specifically explore natural language descriptions as semantic signals, demonstrating their power in narrating nuanced human activities and intentions. To facilitate this connection, we propose a novel task for dense video captioning tailored for videos of egocentric procedural activities. This task aims to generate detailed, temporally coherent natural language descriptions that explain not only the hand's actions but also the context and purpose.

1.6 Dissertation Outline

The remainder of this dissertation is organized as follows and summarized in Fig. 1.4.

Chapter 2 provides a comprehensive survey of the state-of-the-art in 3D hand capture, annotation, and learning methods, discussing the challenges and advancements in this field. This chapter integrates insights from our work titled **"Efficient Annotation and Learning for 3D Hand Pose Estimation: A Survey [251]" (IJCV 2023)**.

Chapter 3 details our contributions to egocentric hand pose estimation under

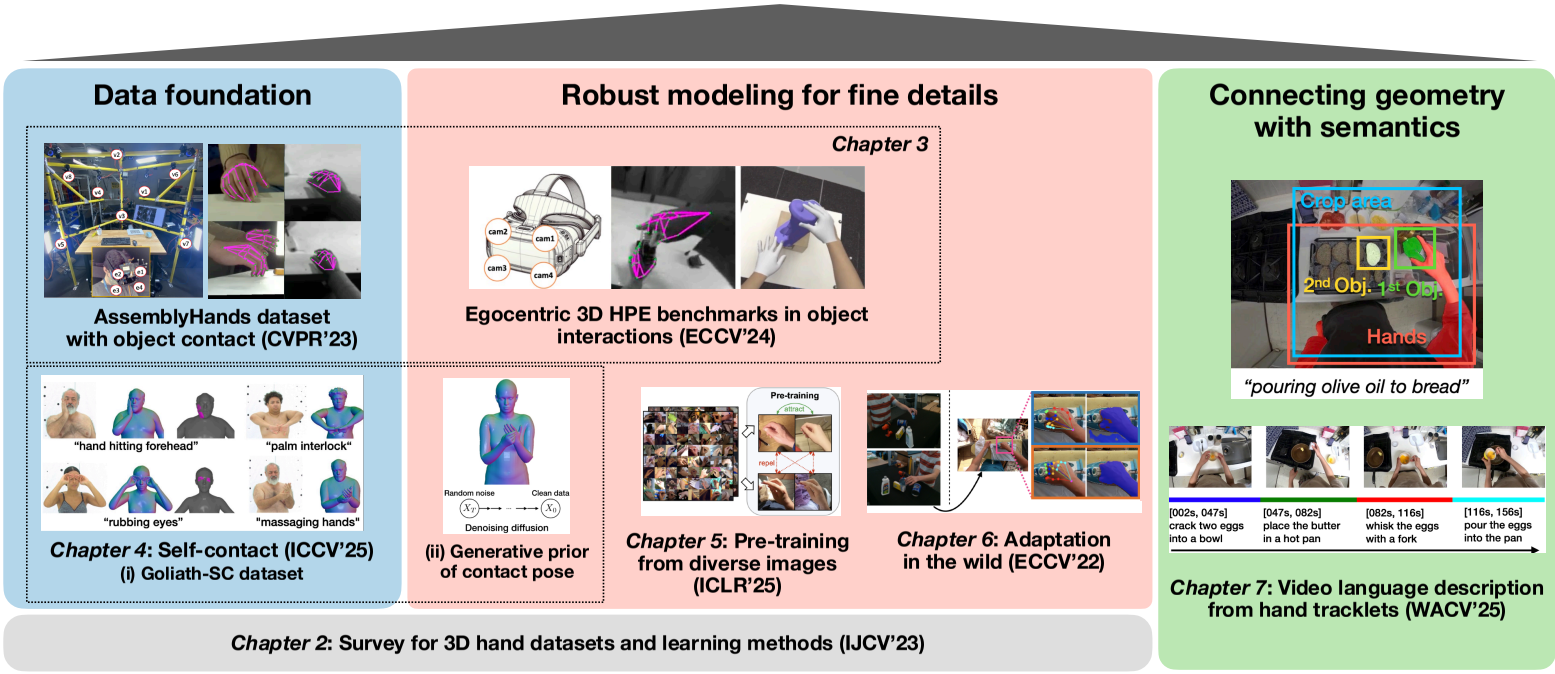


Figure 1.4: **Overview of the dissertation.** Each research work is composed based on the dissertation's goal and the structed three pillars. Corresponding chapters and publications are highlighted in the figure.

complex object interactions. This chapter presents extensive benchmark analysis featuring hand-object contact, underpinning both the Data Foundation (Pillar 1) through new data insights and Robust Modeling (Pillar 2) with improved estimation techniques. This chapter is organized based on the publication of **"Benchmarks and Challenges in Pose Estimation for Egocentric Hand Interactions with Objects [89]" (ECCV 2024)**. This chapter's findings are built upon the preceding publication of **"AssemblyHands: Towards Egocentric Activity Understanding via 3D Hand Pose Estimation [252]" (CVPR 2023)**.

Chapter 4 introduces a new, large-scale benchmark specifically designed for hand self-contact, alongside our generative pose modeling to address the unique occlusions and ambiguities in self-contact. This work further strengthens the Data Foundation (Pillar 1) and Robust Modeling (Pillar 2), based on the publication titled **"Generative Modeling of Shape-Dependent Self-Contact Human Poses [254]" (ICCV 2025)**.

Chapter 5 focuses on advancing Robust Modeling (Pillar 2) by presenting a novel approach for 3D hand pose pre-training from diverse, in-the-wild images. This methodology aims to build more generalizable and robust hand models. This chapter is based on our recent work titled **"SiMHand: Mining Similar Hands for Large-Scale 3D Hand Pose Pre-training [194]" (ICLR 2025)**.

Chapter 6 extends our efforts in Robust Modeling (Pillar 2) by proposing domain adaptation of hand state estimation techniques designed for in-the-wild environments. This chapter details methods to bridge performance gaps across varied recording conditions, building upon our work titled **"Domain Adaptive Hand Keypoint and Pixel Localization in the Wild [255]" (ECCV 2022)**.

Chapter 7 addresses the Connecting Geometry and Semantics (Pillar 3). It introduces a novel task for dense video captioning derived from hand-object tracklets, demonstrating how geometric information can be leveraged for rich semantic interpretation, published as **"Exo2EgoDVC: Dense Video Captioning of Egocentric Procedural Activities Using Web Instructional Videos [257]" (WACV 2025)**.

Chapter 8 concludes this dissertation. It summarizes the main contributions, discusses the broader implications of our findings, outlines current limitations, and proposes promising directions for future research in the field of visual hands understanding.

Chapter 2

Survey for 3D Hand Capture, Annotation, and Learning

In this chapter, we present a systematic review of 3D hand pose estimation from the perspective of efficient annotation and learning. 3D hand pose estimation has been an important research area owing to its potential to enable various applications, such as video understanding, AR/VR, and robotics. However, the performance of models is tied to the quality and quantity of annotated 3D hand poses. Under the status quo, acquiring such annotated 3D hand poses is challenging, *e.g.*, due to the difficulty of 3D annotation and the presence of occlusion. To reveal this problem, we review the pros and cons of existing annotation methods classified as manual, synthetic-model-based, hand-sensor-based, and computational approaches. Additionally, we examine methods for learning 3D hand poses when annotated data are scarce, including self-supervised pretraining, semi-supervised learning, and domain adaptation. Based on the study of efficient annotation and learning, we further discuss limitations and possible future directions in this field.

2.1 Introduction

The acquisition of 3D hand pose annotations [1] has presented a significant challenge in the study of 3D hand pose estimation. This makes it difficult to con-

[1] We denote 3D pose as the 3D keypoint coordinates of hand joints, $\mathrm{P}^{3\mathrm{D}} \in \mathbb{R}^{J\times 3}$ where J is the number of joints.

struct large training datasets and develop models for various target applications, such as hand-object interaction analysis [32, 116], pose-based action recognition [148, 303, 330], augmented and virtual reality [119, 190, 362], and robot learning from human demonstration [75, 120, 214, 276]. In these application scenarios, we must consider methods for annotating hand data, and select an appropriate learning method according to the amount and quality of the annotations. However, there is currently no established methodology that can give annotations efficiently and learn even from imperfect annotations. This motivates us to review methods for building training datasets and developing models in the presence of these challenges in the annotation process.

During the annotations, we encounter several obstacles including the difficulty of 3D measurement, occlusion, and dataset bias. As for the first obstacle, annotating 3D points from a single RGB image is an ill-posed problem. While annotation methods using hand markers, depth sensors, or multi-view cameras can provide 3D positional labels, these setups require a controlled environment, which limits available scenarios. As for the second obstacle, occlusion hinders annotators from accurately localizing the positions of hand joints. As for the third obstacle, annotated data are biased to a specific condition constrained by the annotation method. For instance, annotation methods based on hand markers or multi-view setups are usually installed in laboratory settings, resulting in a bias toward a limited variety of backgrounds and interacting objects.

Given such challenges in annotation, we conduct a systematic review of the literature on 3D hand pose estimation from two distinct perspectives: *efficient annotation* and *efficient learning* (see Fig. 2.1). The former view highlights how existing methods assign reasonable annotations in a cost-effective way, covering a range of topics: the availability and quality of annotations and the limitations when deploying the annotation methods. The latter view focuses on how models can be developed in scenarios where annotation setups cannot be implemented or available annotations are insufficient.

In contrast to existing surveys on network architecture and modeling [51, 82, 172, 179, 201], our survey delves into another fundamental direction that arises from the annotation issues, namely, dataset construction with cost-effective annotation and model development with limited resources. In particular, our survey includes benchmarks, datasets, image capture setups, automatic annotation, learn-

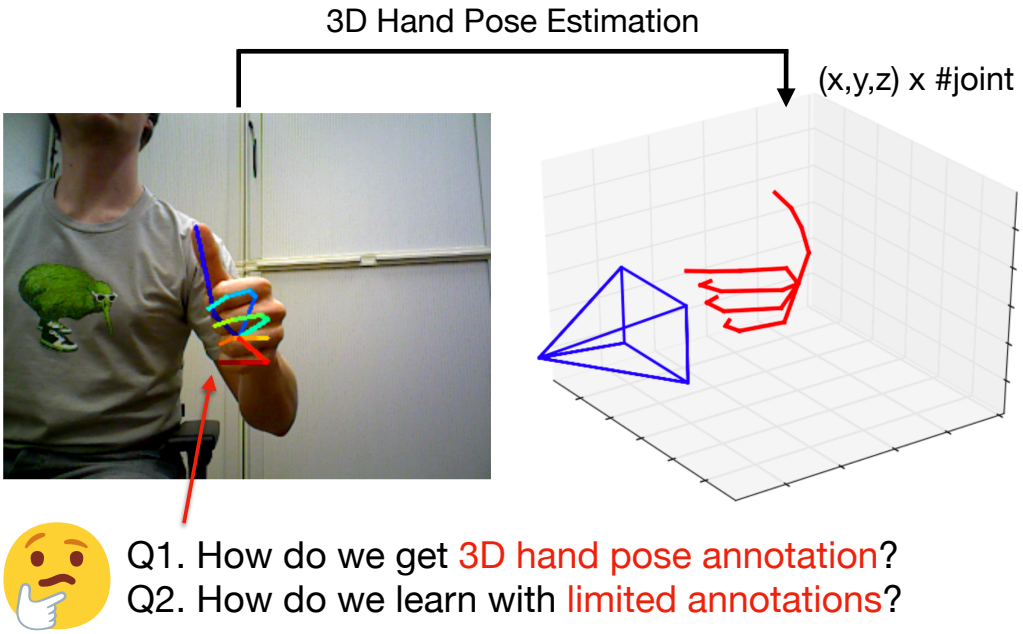


Figure 2.1: **Structure of our survey**. Our study on 3D hand pose estimation is organized from two aspects: (i) obtaining 3D hand pose annotation and (ii) learning even with a limited amount of annotated data. These two issues will be considered in the scenarios of practical applications where we work on dataset construction and model development with limited resources. The figure is adapted from [420].

ing with limited labels, and transfer learning. Finally, we discuss potential future directions of this field beyond the current state of the art.

For the study of annotation, we categorize existing methods into manual [49, 235, 320], synthetic-model-based [55, 127, 233, 235, 420], hand-marker-based [103, 324, 392], and computational approaches [116, 169, 171, 230, 309, 421]. While manual annotation requires querying human annotators, hand markers automate the annotation process by tracking sensors attached to a hand. Synthetic methods utilize computer graphics engines to render plausible hand images with precise keypoint coordinates. Computational methods assign labels by fitting a hand template model to the observed data or using multi-view geometry. We find these annotation methods have their own constraints, such as the necessity of human effort, the sim-to-real gap, the changes in hand appearance, and the limited porta-

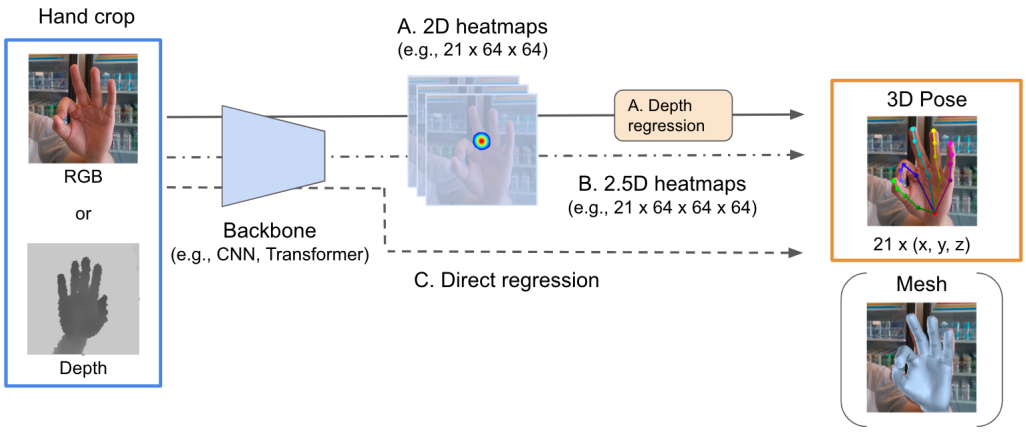


Figure 2.2: **Formulation and modeling of single-view 3D hand pose estimation.** For input, we use either RGB or depth images cropped to the hand region. The model learns to produce a 3D hand pose defined by 3D coordinates. Some works additionally estimate hand shape using a 3D hand template model. For modeling, there are three major designs; (A) 2D heatmap regression and depth regression, (B) extended three-dimensional heatmap regression called 2.5D heatmaps, and (C) direct regression of 3D coordinates.

bility of the camera setups. Thus, these annotation methods may not always be adopted for every application.

Due to the problems and constraints of each annotation method, we need to consider how to develop models even when we do not have enough annotations. Therefore, learning with a small amount of labels is another important topic. For learning from limited annotated data, leveraging a large pool of unlabeled hand images as well as labeled images is a primary interest, *e.g.*, in self-supervised pretraining, semi-supervised learning, and domain adaptation. Self-supervised pretraining encourages the hand pose estimator to learn from unlabeled hand images, so it enables building a strong feature extractor before performing supervised learning. While semi-supervised learning trains the estimator with labeled and unlabeled hand images collected from the same environment, domain adaptation further solves the so-called problem of domain gap between the two image sets, *e.g.*, the difference between synthetic data and real data.

The rest of this survey is organized as follows. In Sec. 2.2, we introduce the formulation and modeling of 3D hand pose estimation. In Sec. 2.3, we present

open challenges in the construction of hand pose datasets involving depth measurement, occlusion, and dataset bias. In Sec. 2.4, we cover existing methods of 3D hand pose annotation, namely manual, synthetic-model-based, hand-marker-based, and computational approaches. In Sec. 2.5, we provide learning methods from a limited amount of annotated data, namely self-supervised pretraining, semi-supervised learning, and domain adaptation. In Sec. 2.6, we finally show promising future directions of 3D hand pose estimation.

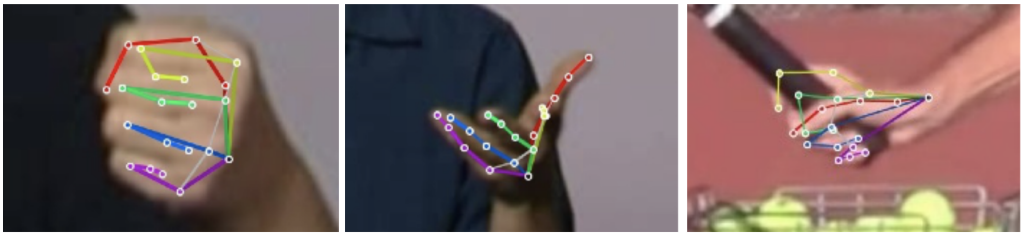
2.2 Overview of 3D Hand Pose Estimation

Task setting: As shown in Fig. 2.2, 3D hand pose estimation is typically formulated as the estimation from a monocular RGB/depth image [86, 147, 390]. The output is parameterized by the hand joint positions with 14, 16, or 21 keypoints, which are introduced in [333], [325], and [274], respectively. The dense representation of 21 hand joints [2] has been popularly used as it contains more precise information about hand structure. For a single RGB image in which depth and scale are ambiguous, the 3D coordinates of the hand joint relative to the hand root are estimated from a scale-normalized hand image [40, 104, 420]. Recent works additionally estimate hand shape by regressing 3D hand pose and shape parameters together [32, 104, 234, 410]. In evaluation, produced prediction is compared with ground truth, *e.g.*, in the space of world or image coordinates. These two metrics are often used: mean per joint position error (MPJPE) in millimeters, and area under curve of percentage of correct keypoints (PCK-AUC).

Modeling: Classic methods estimate a hand pose by finding the closest sample from a large set of hand poses, *e.g.*, synthetic hand pose sets. Some works formulate the task as nearest neighbor search [289, 292] while others solve pose classification given predefined hand pose classes and a SVM classifier [288, 290, 321].

Recent studies have adopted an end-to-end training manner where models learn the correspondence between the input image and its label of the 3D hand pose. Standard single-view methods from an RGB image [40, 104, 420] consist of (A) the estimation of 2D hand poses by heatmap regression and depth regression for each 2D keypoint (see Fig. 2.2). The 2D keypoints are learned by

[2] Five end keypoints are fingertips, not strictly called joints.



(a) Articulation (b) Viewpoint (c) Object

Figure 2.3: **Difficulty of hand pose annotation in a single RGB image [309].** Occlusion of hand joints is caused by (a) articulation, (b) viewpoint bias, and (c) grasping objects.

optimizing heatmaps centered on each 2D hand joint position. An additional regression network predicts the depth distance of detected 2D hand keypoints. Other works use (B) extended 2.5D heatmap regression with a depth-wise heatmap in addition to the 2D heatmaps [149, 230], so it does not require a depth regression branch. Depth-based hand pose estimation also utilizes such heatmap regression [145, 286, 366]. Instead of the heatmap training, other methods learn to (C) directly regress keypoint coordinates [299, 319].

For the architecture of the backbone network, CNNs (*e.g.*, ResNet [131]) are a basic choice while recent Transformer-based methods have been proposed [117, 144]. To generate feasible hand poses, regularization is a key trick in correcting predicted 3D hand poses. Based on the anatomical study of hands, bio-mechanical constraints are imposed to limit predicted bone lengths and joint angles [65, 200, 317].

2.3 Challenges in Dataset Construction

Task formulation and algorithms for estimating 3D hand poses are outlined in Sec. 2.2. During training, it is necessary to build a large amount of training data with diverse hand poses, viewpoints, and backgrounds. However, obtaining such massive hand data with accurate annotations has been challenging for the following reasons.

Difficulty of 3D annotation: Annotating the 3D position of hand joints from a single RGB image is inherently impossible without any prior information or additional sensors due to an ill-posed condition. To assign accurate hand pose labels, hand-marker-based annotation using magnetic sensors [103, 360, 392], motion capture systems [225, 301, 324], or hand gloves [27, 107, 353] has been studied. These sensors can provide 6-DoF information (*i.e.*, location and orientation) of attached markers and enable us to calculate the coordinates of full hand joints from the tracked markers. However, their setups are expensive and need good calibration, which constrains available scenarios.

On the contrary, depth sensors (*e.g.*, RealSense) or multi-view camera studios [49, 116, 230, 309, 421] make it possible to obtain depth information near hand regions. Given 2D keypoints for an image, these setups enable annotation of 3D hand poses by measuring the depth distance at each 2D keypoint. However, these annotation methods do not always produce satisfactory 3D annotations, *e.g.*, due to an occlusion problem (detailed in the next section). In addition, depth images are significantly affected by the sensor noise, such as unknown depth values in some regions and ghost shadows around object boundaries [367]. Due to the limited depth distance that depth cameras can capture, the depth measurement becomes inaccurate when the hands are far from the sensor.

Occlusion: Hand images often contain complex occlusions that distract human annotators from localizing hand keypoints. Examples of possible occlusions are shown in Fig. 2.3. In figure (a), articulation causes a self-occlusion that makes some hand joints (*e.g.*, fingertips) invisible due to the overlap with the other parts of the hand. In figure (b), such self-occlusion depends on a specific camera viewpoint. In figure (c), hand-held objects induce occlusion that hides the hand joints by the object during the interaction.

To address this issue, hand-marker-based tracking [103, 324, 360, 392] and multi-view camera studios [49, 116, 230, 309, 421] have been studied. The hand markers offer 6-DoF information during these occlusions, so the hand-maker-based annotation is robust to the occlusion. For multi-camera settings, the effect of occlusion can be reduced when many cameras are densely arranged.

Dataset bias: While hands are a common entity in various image capture settings, the category of objects, including hand-held objects (*i.e.*, foregrounds) and backgrounds, is potentially diverse. In order to improve the generalization ability

Figure 2.4: **Example of major data collection setups.** The synthetic image on the left (ObMan [127]) can be generated inexpensively, but they exhibit unrealistic hand texture. The hand markers on the middle (FPHA [103]) enable automatic tracking of hand joints, although the markers distort the appearance of hands. The in-lab setup on the right (DexYCB [49]) uses a black background to make it easier to recognize hands and objects, but it limits data variation in environments.

of hand pose estimators, hand images must be annotated under various imaging conditions (*e.g.*, lighting, viewpoints, hand poses, and backgrounds). However, it is challenging to create such large and diverse datasets nowadays due to the aforementioned problems. Rather, existing hand pose datasets exhibit a bias to a particular imaging condition constrained by the annotation method.

As shown in Fig. 2.4, generating data using synthetic models [55, 127, 233, 235, 420] is cost-effective, but it creates unrealistic hand texture [256]. Although the hand-marker-based annotation [103, 324, 360, 392] can automatically track the hand joints from the information of hand sensors, the sensors distort the hand appearance and hinder the natural hand movement. In-lab data acquired by multi-camera setups [49, 116, 230, 309, 421] make the annotation easier because they can reduce the occlusion effect. However, the variations in environments (*e.g.*, backgrounds and interacting objects) are limited because the setups are not easily portable.

2.4 Annotation Methods

Given the above challenges concerning the construction of hand pose datasets, we review existing 3D hand pose datasets in terms of annotation design. As shown in

Annotation	Dataset	Year	Modality	Resolution	#Frame	#Subj./ #Obj.	#View	Motion	Obj. pose	Sensor/Engine
Manual	MSRA [274]	2014	Depth	320×240	2K	6 / -	1 (3rd)	×	×	Creative Senz3D
	Dexter+Object [320]	2016	RGB-D	640×480	3K	2 / 2	1 (3rd)	✓	✓	Creative Senz3D
	EgoDexter [235]	2017	RGB-D	640×480	3K	4 / -	1 (1st)	✓	×	RealSense SR300
Synthetic	SynthHands [235]	2017	RGB-D	640×480	220K	- / 7	5 (1st)	×	×	Unity
	RHD [420]	2017	RGB	320×320	43K	20 / -	1 (3rd)	×	×	Unity
	GANerated [233]	2018	RGB	256×256	331K	- / -	1 (3rd)	×	×	Unity
	ObMan [127]	2019	RGB-D	256×256	154K	20 / 3K	1 (3rd)	×	✓	Blender/GraspIt [223]
	MVHM [55]	2021	RGB-D	256×256	320K	- / -	8 (3rd)	×	×	Blender
Hand marker	BigHand2.2M [392]	2017	Depth	640×480	2,200K	10 / -	1 (1st + 3rd)	✓	×	NDI trakSTAR
	FPHA [103]	2018	RGB-D	640×480	105K	6 / 4	1 (1st)	✓	✓	NDI trakSTAR
	GRAB [324]	2020	MoCap	-	1,624K	10 / 51	-	✓	✓	VICON Vantage 16
Computational										
Model fitting	ICVL [325]	2014	Depth	320×240	180K	10 / -	1 (3rd)	×	×	Creative Senz3D
	NYU [333]	2014	RGB-D	640×480	81K	2 / -	1 (3rd)	×	×	PrimeSense
	FreiHAND [421]	2019	RGB	224×224	37K	32 / 27	8 (3rd)	×	×	Basler ace
	YouTube3DHands [169]	2020	RGB	640×480	47K	(109) / -	1 (3rd)	✓	×	-
	HO-3D [116]	2020	RGB-D	640×480	103K	10 / 10	1-5 (3rd)	✓	✓	
	DexYCB [49]	2021	RGB-D	640×480	582K	10 / 20	8 (3rd)	✓	✓	RealSense D415
	H2O [171]	2021	RGB-D	1280×720	571K	4 / 10	5 (1st & 3rd)	✓	✓	Azure Kinect
Triangulation	Panoptic Studio [309]	2017	RGB	1920×1080	15K	- / -	31 (3rd)	✓	×	HD camera
	InterHand2.6M [230]	2020	RGB	512×334	2,590K	27 / -	80-140 (3rd)	✓	×	HD camera
	AssemblyHands [252]	2023	RGB/Mono	636×480	3,030K	34 / -	12 (1st & 3rd)	✓	×	HD camera/Meta Quest

Table 2.1: **Taxonomy of methods for annotating 3D hand poses**. We categorize the annotation methods as manual, synthetic-model-based, hand-marker-based, and computational annotation.

Tab. 2.1, we categorize the annotation methods as manual, synthetic-model-based, hand-marker-based, and computational approaches. We then study the pros and cons of each annotation method in Tab. 2.2.

2.4.1 Manual annotation

MSRA [274], Dexter+Object [320], and EgoDexter [235] manually annotate 2D hand keypoints on the depth images and determine the depth distance from the depth value of the images on the 2D point. This method enables assigning reasonable annotations of 3D coordinates (*i.e.*, 2D position and depth) when hand joints are fully visible.

However, it is not extensively available according to the number of frames due to the high annotation cost. In addition, since it is not robust for occluded keypoints, this approach only allows fingertip annotation, instead of full hand joints. For these limitations, these datasets provide a small amount of data ($\approx$ 3K images) used for evaluation only. Additionally, these single-view datasets can produce view-dependent annotation errors because a single-depth camera captures the distance to the hand skin surface, not the true joint position. To reduce such unavoidable errors, subsequent annotation methods based on multi-camera setups provide further accurate annotations (see Sec. 2.4.4).

2.4.2 Synthetic-model-based annotation

To acquire large-scale hand images and labels, synthetic methods based on synthetic hand and full-body models [206, 290, 293, 346] have been proposed. SynthHands [235] and RHD [420] render synthetic hand images with randomized real backgrounds from either a first- or third-person view. MVHM [55] generates multi-view synthetic hand data rendered from eight viewpoints. These datasets have succeeded in providing accurate hand keypoint labels on a large scale. Although they can generate various background patterns inexpensively, the lighting and texture of hands are not well simulated, and the simulation of hand-object interaction is not considered in the data generation process.

To handle these issues, GANerated [233] utilizes GAN-based image translation to stylize synthetic hands more realistically. Furthermore, ObMan [127] simulates the hand-object interaction in data generation using a hand grasp simulator

Annotation	Pros	Cons
Manual	Reasonable accuracy	Labor intensive
		Hard to address occlusion
Synthetic	Large scale	Sim-to-real gap
	High diveristy	Hard to simulate motion
	Low cost	
Hand marker	Robust to occlusion	Requires special sensors
	Low annotation cost	Changes visual modality
		Prevents natural motion
Computational	Natural motion	Lacks diversity
	Low annotation cost	Hard to evaluate quality
		Needs multi-camera setups

Table 2.2: **Pros and cons of each annotation approach.**

(Graspit [223]) with known 3D object models (ShapeNet [48]). Ohkawa *et al*. proposed foreground-aware image stylization to convert the simulation texture in the ObMan data to a more realistic one while separating the hand regions and backgrounds [256]. Corona *et al*. attempted to synthesize more natural hand grasps with affordance classification and the refinement of fingertip locations [76]. However, the ObMan data only provide static hand images with hand-held objects, not hand motion. The hand motion simulation while approaching the object remains an open problem.

2.4.3 Hand-marker-based annotation

As shown in Fig. 2.5, hand-marker-based annotation automatically tracks attached hand markers and then calculates the coordinates of hand joints. Initially, Wetzler *et al*. attached magnetic sensors to fingertips that provide 6-DoF information of the markers [360]. While this scheme can annotate fingertips only, recent datasets, BigHand2.2M [392] and FPHA [103], use these sensors to offer the annotation of the full 21 hand joints. Fig. 2.6 shows how to compute the joint positions given six magnetic sensors. It uses inverse kinematics to infer all 21 hand joints, which fits a hand skeleton with the constraints of the maker positions and

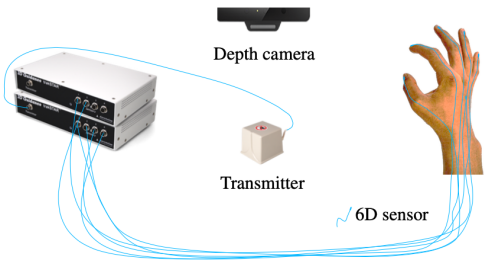


Figure 2.5: **Hand marker setup [392].**

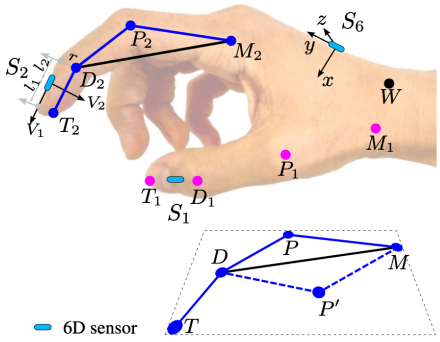


Figure 2.6: **Calculation of joint positions from tracked markers [392].**

user-specific bone length manually measured beforehand.

However, these sensors obstruct natural hand movement and distort the appearance of the hand. Due to the changes in hand appearance, these datasets have been proposed for the benchmark of depth-based estimation, not the RGB-based task. On the contrary, GRAB [324] is built with a motion capture system for human hands and body, but it does not possess visual modality, *e.g.*, RGB images.

2.4.4 Computational annotation

Computational annotation is categorized into two major approaches: *hand model fitting* and *triangulation*. Unlike hand-marker-based annotation, these methods can capture natural hand motion without attaching hand markers.

Model fitting (depth): Early works of computational annotation utilize *model fitting* on depth images [147, 390]. Since a depth image provides 3D structural information, their works fit a 3D hand model, from which joint positions can be obtained, to the depth image. ICVL [325] fits a convex rigid body model by solving a linear complementary problem with physical constraints [218]. NYU [333] uses a hand model defined by spheres and cylinders and formulates the model fitting as a kind of particle swarm optimization [258, 259]. The use of other cues for the model fitting is also studied [15, 207], such as edges, optical flow, shading, and collisions. Sharp *et al*. paint hands to obtain hand part labels by color segmentation on RGB images and the proxy cue of hand parts further helps the depth-based model fitting [306].

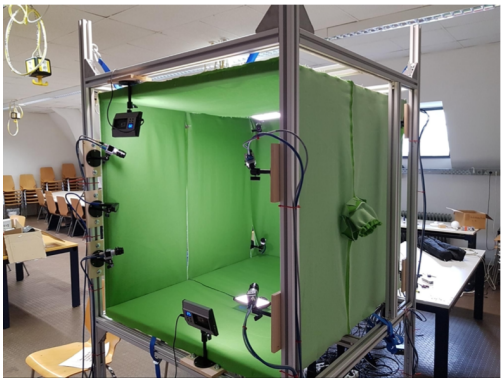


Figure 2.7: **Multi-camera setup [421].**

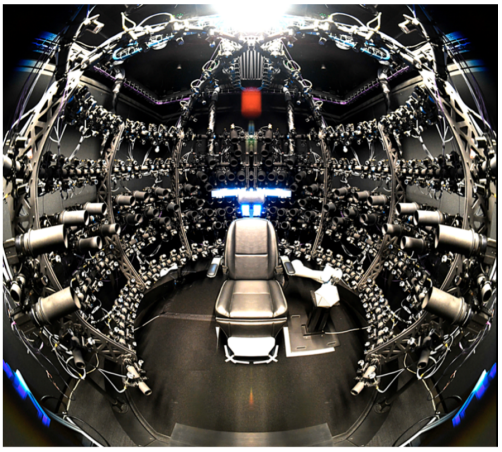


Figure 2.8: **Many-camera setup [363] from [230].**

Using these depth datasets, several more accurate labeling methods have been proposed. Rogez *et al*. gave manual annotation to a few joints and searched the closest 3D pose from a pool of synthetic hand pose data [290]. Oberweger *et al*. considered model fitting with temporal coherence [248]. This method selects reference frames from a depth video and asks annotators for manual labeling. Model fitting is done separately for annotated reference frames and unlabeled non-reference frames. Finally, all sequential poses are optimized to satisfy temporal smoothness.

Triangulation (RGB): For the annotation of RGB images, a multi-camera studio is often used to compute 3D points by multi-view geometry, *i.e*., *triangulation* (see Fig. 2.7). Panoptic Studio [309] and InterHand2.6M [230] triangulate a 3D hand pose from multiple 2D hand keypoints provided by an open source library, OpenPose [135], or human annotators. The generated 3D hand pose is reprojected onto the image planes of other cameras to annotate hand images with novel viewpoints. This multi-view annotation scheme is beneficial when many cameras are installed (see Fig. 2.8). For instance, the InterHand2.6M manually annotates keypoints from 6 views and reprojects the triangulated points to the other many views (100+). This setup can produce over 100 training images for every single annotation. The InterHand2.6M has million-scale training data.

This point-level triangulation method works quite well when many cameras

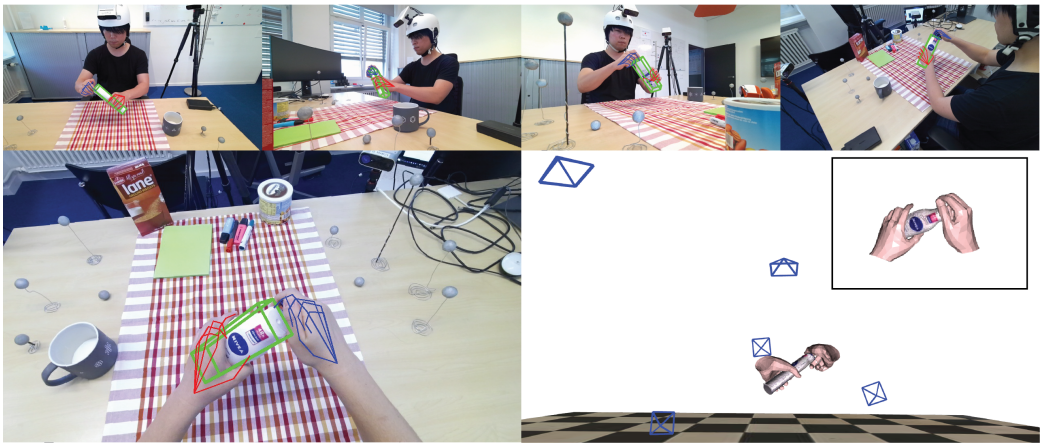


Figure 2.9: **Synchronized multi-camera setup with first-person and third-person cameras [171].**

(30+) are arranged [230, 309]. However, the AssemblyHands setup [252] has only eight static cameras, and then the predicted 2D keypoints to be triangulated tend to be suboptimal due to hand-object occlusion during the assembly task. To improve the accuracy of triangulation in such sparse camera settings, Ohkawa *et al*. adopt multi-view aggregation of encoded features by the 2D keypoint detector and compute 3D coordinates from constructed 3D volumetric features [23, 150, 252, 421]. This feature-level triangulation provides better accuracy than the point-level method, achieving an average keypoint error of 4.20 mm, which is 85% lower than the error of the original annotations in Assembly101 [303].

Model fitting (RGB): Model fitting is also used in RGB-based pose annotation. FreiHAND [419, 421] utilizes a 3D hand template (MANO [293]) fitting to multi-view hand images with sparse 2D keypoint annotation. The dataset increases the variation of training images by randomly synthesizing the background and using captured real hands as the foreground. YouTube3DHands [169] uses the MANO model fitting to estimated 2D hand poses in YouTube videos. HO-3D [116], DexYCB [49], and H2O [171] jointly annotate 3D hand and object poses to facilitate a better understanding of hand-object interaction. Using estimated or manually annotated 2D keypoints, their datasets fit the MANO model and 3D object models to the hand images with objects.

While most methods capture hands from static third-person cameras, H2O and AssemblyHands install first-person cameras that are synchronized with static

third-person cameras (see Fig. 2.9). With camera calibration and head-mounted camera tracking, such camera systems can offer 3D hand pose annotations for first-person images by projecting annotated keypoints from third-person cameras onto first-person image planes. This reduces the cost of annotating first-person images, which is considered expensive because the image distribution changes drastically over time and the hands are sometimes out of view.

These computational methods can generate labels with little human effort, although the camera system itself is costly. However, assessing the quality of the labels is still difficult. In fact, the annotation quality depends on the number of cameras and their arrangement, the accuracy of hand detection and the estimation of 2D hand poses, and the performance of triangulation and fitting algorithms.

2.5 Learning with Limited Labels

As explained in Sec. 2.4, existing annotation methods have certain pros and cons. Since perfect annotation in terms of amount and quality cannot be assumed, training 3D hand pose estimators with limited annotated data is another important study. Accordingly, we introduce learning methods using unlabeled data, namely self-supervised pretraining, semi-supervised learning, and domain adaptation.

2.5.1 Self-supervised pretraining and learning

Self-supervised pretraining aims to utilize massive unlabeled hand images and build an improved encoder network before supervised learning with labeled images. As shown in Fig. 2.10, recent works [316, 419] first pretrain an encoder network that extracts image features by using contrastive learning (*e.g.*, MoCo [130] and SimCLR [59]) and then fine-tune the whole network in a supervised manner. The core idea of contrastive learning is to push a pair of *similar* instances closer together in an embedding space while unrelated instances are pushed apart. This approach focuses on how to define the *similarity* of hand images and how to design embedding techniques. Spurr *et al*. proposed to geometrically align two features generated from differently augmented instances [316]. Zimmermann *et al*. found that multi-view images representing the same hand pose can be effective pair supervision [419].

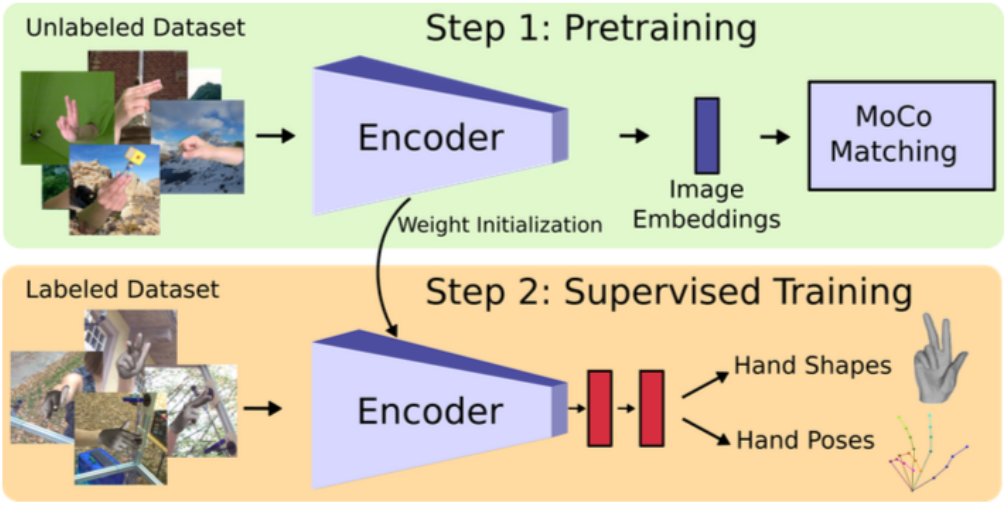


Figure 2.10: **Self-supervised pretraining of 3D hand pose estimation [419].** The pretraining phase (step 1) aims to construct an improved encoder network by using many unlabeled data before supervised learning (step 2). The work uses MoCo [130] as a method of self-supervised learning.

Other works utilize the scheme of self-supervised learning that solves an auxiliary task, instead of the target task of hand pose estimation. Given the prediction on an unlabeled depth image, [249, 348] render a synthetic depth image and penalize the matching between the input image and the one generated from the prediction. This auxiliary loss by image synthesis is informative even when annotations are scarce.

2.5.2 Semi-supervised learning

As shown in Fig. 2.11, semi-supervised learning is used to learn from small labeled data and large unlabeled data simultaneously. Liu *et al*. proposed a pseudo-labeling method that learns unlabeled instances with pseudo-ground-truth given from the model's prediction [200]. This pseudo-label training is applied only when its prediction satisfies spatial and temporal constraints. The spatial constraints check the correspondence of a 2D hand pose and the 2D pose projected from 3D hand pose prediction. In addition, they include a constraint based on bio-mechanical feasibility, such as bone lengths and joint angles. The temporal

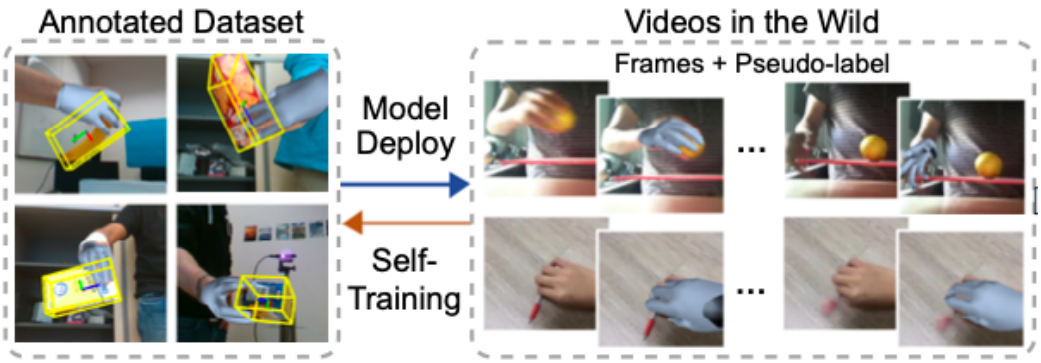


Figure 2.11: **Semi-supervised learning of 3D hand pose estimation [200].** The model is trained jointly on annotated data and unlabeled data with pseudo-labels.

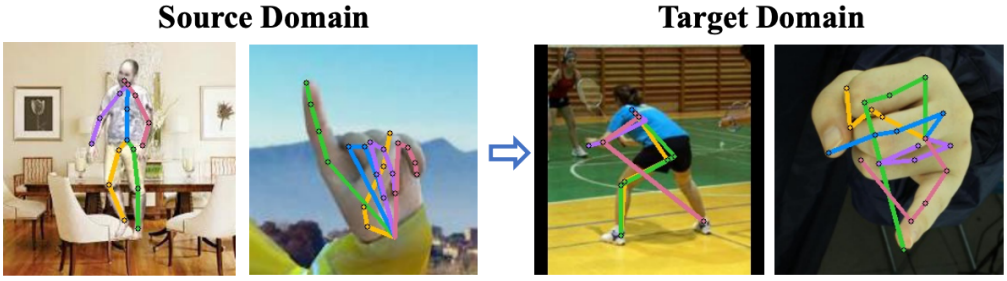


Figure 2.12: **Poor generalization to an unknown domain [151].** The models trained on synthetic images (source) exhibit a limited capacity for inferring poses on real images (target).

constraints indicate the smoothness of hand pose and mesh predictions over time.

Yang *et al*. proposed the combination of pseudo-labeling and consistency training [377]. In pseudo-labeling, the generated pseudo-labels are corrected by fitting the hand template model. In addition, the work enforces consistency losses between the predictions of differently augmented instances and between the modalities of 2D hand poses and hand masks.

Spurr *et al*. applied adversarial training to a sequence of predicted hand poses [318]. The encoder network is expected to be improved by fooling a discriminator that distinguishes between plausible and invalid hand poses.

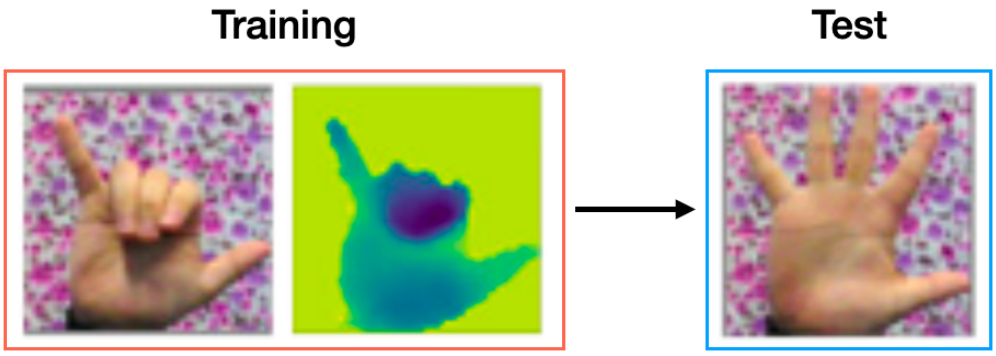


Figure 2.13: **Example of modality transfer**. During training, RGB and depth images are accessible and RGB images are given in the test phase. The training aims to utilize the support of depth information to improve RGB-based hand pose estimation.

2.5.3 Domain adaptation

Domain adaptation aims to improve model performance on target data by learning from labeled source data and target data with limited labels. This study has addressed two types of underlying domain gaps: *between different datasets* and *between different modalities*.

The former problem *between different datasets* is a common domain adaptation problem where the source and target data are sampled from two datasets with different image statistics, *e.g.*, sim-to-real adaptation [151, 326] (see Fig. 2.12). The model has access to readily available synthetic images with labels and target real images without labels. The latter problem *between different modalities* is characterized as *modality transfer* where the source and target data represent the same scene, but the modalities are different, *e.g.*, depth *vs.* RGB (see Fig. 2.13). This aims to utilize information-rich source data, *e.g.*, depth images with 3D structural information, for inferring easily available target data (*e.g.*, RGB images).

To reduce the gap between the two datasets, two major approaches have been proposed: *generative methods* and *adversarial methods*. In generative methods, Qi *et al.* proposed an image translation method to alter the synthetic textures to realistic ones and train a model on generated real-like synthetic data [273].

Adversarial methods enforce matching two domains' features so that the feature extractor can encode features even from the target domain. However, in ad-

dition to the domain gap in an input space (*e.g.*, the difference in backgrounds), the gap in a label space also exists in this task, which is not assumed in typical adversarial methods [102, 337]. Zhang *et al.* developed a feature matching method based on Wasserstein distance and proposed adaptive weighting to enable matching only for features related to hand characteristics, except for label information [403]. Jiang *et al.* utilized an adversarial regressor and optimized the domain disparity by a minimax game [151]. Such minimax of disparity is effective in domain adaptation of regression tasks, including hand pose estimation.

As for the *modality transfer* problem, Yuan *et al.* and Rad *et al.* attempted to use depth images as the auxiliary information during training and test the model on RGB images [279, 391]. They observed that learned features from depth images could support RGB-based hand pose estimation. Park *et al.* transferred the knowledge from depth images to infrared (IR) images that have less motion blur [264]. Their training is facilitated by matching two features from paired images, *e.g.*, (RGB, depth) and (depth, IR). Baek *et al.* newly defined the domain of hand-only images where a hand-held object is removed. The work translates hand-object images to hand-only images by using GAN and mesh renderer [14]. Given a hand-object image with an unknown object, this method can generate hand-only images, from which hand pose estimation is more tractable.

2.6 Future Directions

2.6.1 Flexible camera systems

We believe that hand image capture will feature more flexible camera systems, such as using first-person cameras. To reduce the occlusion effect without the need for hand markers, recently published hand datasets have been acquired by multi-camera setups, *e.g.*, DexYCB [49], InterHand2.6M [230], and FreiHAND [421]. These setups are static and not suitable for capturing dynamic user behavior. To address this, a *first-person camera* attached to the user's head or body is useful because it mostly captures close-up hands even when the user moves around. However, as shown in Tab. 2.1, existing first-person benchmarks have a very limited variety due to heavy occlusion, motion blur, and a narrow field-of-view.

One promising direction is a joint camera setup with first-person and third-

person cameras, such as H2O [171] and AssemblyHands [252]. This results in flexibly capturing the user's hands from the first-person camera while taking the benefits of multiple third-person cameras (*e.g.*, mitigating the occlusion effect). However, the first-person camera wearer doesn't always have to be alone. Image capture with multiple first-person camera wearers in a static camera setup will advance the analysis of multi-person cooperation and interaction, *e.g.*, game playing and construction with multiple people.

2.6.2 Various types of activities

We believe that increasing the type of activities is an important direction for generalizing models to various situations with hand-object interaction. A major limitation of existing hand datasets is the narrow variation of users' performing tasks and grasping objects. To avoid object occlusion, some works did not capture hand-object interaction [230, 392, 420]. Others [49, 116, 127] used pre-registered 3D object models (*e.g.*, YCB [42]) to simplify in-hand object pose estimation. User action is also very simple in these benchmarks, such as *pick and place*.

From an affordance perspective [124], diversifying the object category will result in increasing hand pose variation. Potential future works will capture goal-oriented and procedural activities that naturally occur in our daily life [77, 111, 303], such as cooking, art and craft, and assembly.

To enable this, we need to develop portable camera systems and robust annotation methods for complex backgrounds and unknown objects. In addition, occurring hand poses are constrained to the context of the activity. Thus, pose estimators conditioned by actions, objects, or textual descriptions of the scene will improve estimation in various activities.

2.6.3 Towards minimal human effort

Secs. 2.4 and 2.5 separately explain efficient annotation and learning. To minimize the effort of human intervention, utilizing findings from both annotation and learning perspectives is one of the promising directions. Feng *et al*. exploited *active learning* that optimizes which unlabeled instance should be annotated and semi-supervised learning that jointly utilizes labeled data and large unlabeled data [92]. However, this method is constrained to triangulation-based 3D pose estimation.

As we mentioned in Sec. 2.4.4, another major computational annotation is *model fitting*; thus, we still need to consider such a collaborative approach in the annotation based on model fitting.

Zimmermann *et al*. also proposed a framework of human-in-loop annotation that inspects the annotation quality manually while updating annotation networks on the inspected annotations [420]. However, this human check will be a bottleneck in large dataset construction. The evaluation of annotation quality on the fly is a necessary technique to scale up the combination of annotation and learning.

2.6.4 Generalization and adaptation

Increasing the generalization ability across different datasets or adapting models to a specific domain is a remaining issue. The bias of existing training datasets hinders the estimators from inferring test images captured under very different imaging conditions. In fact, as reported in [118, 421], models trained on existing hand pose datasets poorly generalize to other datasets. For real-world applications (*e.g*., AR), it is crucial to transfer models from indoor hand datasets to outdoor videos because common multi-camera setups are not available outdoors [255]. Thus, aggregating multiple annotated yet biased datasets for generalization and robustly adapting to very different environments are important future tasks.

2.7 Summary

We presented the survey of 3D hand pose estimation from the standpoint of efficient annotation and learning. We provided a comprehensive overview of this task and modeling, and open challenges during dataset construction. We investigated annotation methods categorized as manual, synthetic-model-based, hand-marker-based, and computational approaches, and examined their respective strengths and weaknesses. In addition, we studied learning methods that can be applied even when annotations are scarce, namely self-supervised pretraining, semi-supervised learning, and domain adaptation. Finally, we discussed potential future advancements in 3D hand pose estimation, including next-generation camera setups, increased object and action variation, jointly optimized annotation and learning techniques, and generalization and adaptation.

Chapter 3

Egocentric Hand Pose Estimation under Object Interactions

We interact with the world with our hands and see it through our own (egocentric) perspective. A holistic 3D understanding of such interactions from egocentric views is important for tasks in robotics, AR/VR, action recognition and motion generation. Accurately reconstructing such interactions in 3D is challenging due to heavy occlusion, viewpoint bias, camera distortion, and motion blur from the head movement. In this chapter, we designed the HANDS23 challenge based on the AssemblyHands and ARCTIC datasets with carefully designed training and testing splits. Based on the results of the top submitted methods and more recent baselines on the leaderboards, we perform a thorough analysis on 3D hand(-object) reconstruction tasks. Our analysis demonstrates the effectiveness of addressing distortion specific to egocentric cameras, adopting high-capacity transformers to learn complex hand-object interactions, and fusing predictions from different views. Our study further reveals challenging scenarios intractable with state-of-the-art methods, such as fast hand motion, object reconstruction from narrow egocentric views, and close contact between two hands and objects. Our efforts will enrich the community's knowledge foundation and facilitate future hand studies on egocentric hand-object interactions.

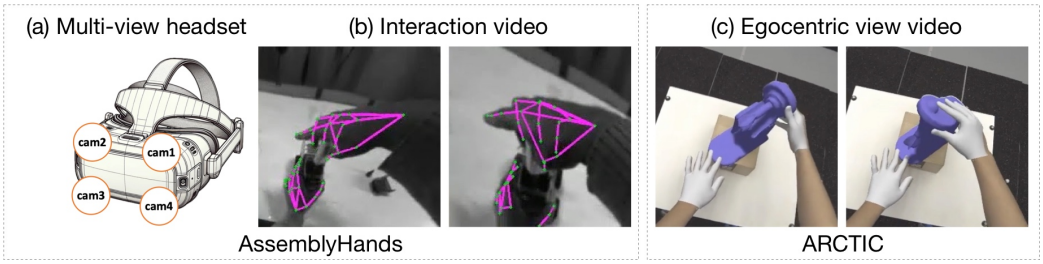


Figure 3.1: **Tasks in HANDS23 based on AssemblyHands and ARCTIC**. In AssemblyHands, from its multi-view headset (a), the goal is to estimate 3D hand poses from images (b); In ARCTIC, given an image, the goal is to estimate the poses of two hands and articulated object surface models (c).

3.1 Introduction

We interact with the world with our hands and see it through our eyes: we wake up and grab our phone to check the time; we use tools when assembling parts of a car; we open the microwave door to heat food, to name a few. An egocentric 3D understanding of our hand interactions with objects will fundamentally impact areas such as robotics grasping [72, 397], augmented and virtual reality [119], action recognition [50, 103, 359] and motion generation [397, 398].

However, it is non-trivial to accurately reconstruct 3D hands and or objects due to its high degree of freedoms [317, 377], ambiguous texture [91], and heavy occlusions. These challenges are intensified in an egocentric view [255, 256], particularly with object interactions, due to significant camera distortion, rapid and varied changes in viewpoint caused by head movements and hand-object occlusion. To better understand these challenges, we introduce a public challenge in conjunction with ICCV 2023 (*i.e.*, HANDS23) based on recent egocentric hand datasets, AssemblyHands [252] and ARCTIC [88] (see Fig. 3.1). These two datasets are large-scale, multi-view, and provide monochrome or RGB egocentric videos of the hands dexterously manipulating objects. Accordingly, we host two tasks: 1) egocentric 3D hand pose estimation from a single-view image based on AssemblyHands, and 2) consistent motion reconstruction based on ARCTIC.

We introduce new HANDS23 methods and recent dataset leaderboard baselines that substantially outperform initial baselines for both tasks, setting new benchmarks for subsequent comparisons on the datasets. The two datasets are sig-

nificantly larger and include a wider variety of bimanual manipulations compared to earlier datasets [49, 116], enabling a more authentic assessment of real-world interactions. With these benchmarks, we thoroughly analyze factors such as viewpoint, action types, hand position, model size, and object variations to determine their impact on 3D hand(-object) reconstruction.

Our findings show the success of addressing the distortion of egocentric cameras with explicit perspective cropping or implicit learning for the distortion bias. In addition, recent high-capacity vision transformers can learn complex hand-object interactions. Adaptive fusion techniques for multi-view predictions further boost performance. We also analyze the remaining challenges that are still difficult to handle with the recent methods, *e.g.*, fast hand motion, object reconstruction from narrow and moving views, and intricate interactions and close contact between two hands and objects.

To summarize, we contribute state-of-the-art baselines and gather the submitted methods for AssemblyHands and ARCTIC to foster future research on egocentric hand-object interactions. Furthermore, we thoroughly analyze the two benchmarks to provide insights for future directions in egocentric hand pose estimation and consistent motion reconstruction.

3.2 Related Work

3D hand pose estimation: Reconstructing 3D hand poses has a long history [86, 251] ever since the important work led by Rehg and Kanade [284]. A large body of research in this area focuses on single-hand reconstruction [32, 40, 49, 62, 84, 91, 98, 104, 123, 149, 198, 233, 309, 316, 317, 319, 340, 402, 418, 420]. For example, while a popular OpenPose library features 2D hand keypoints [309], Zimmermann *et al*. [420] initially extend to estimate 3D hand poses using deep convolutional networks. Ever since the release of InterHand2.6M [230] dataset, the community has increased focus on strongly interacting two hands reconstruction [115, 171, 174, 183, 184, 220, 226, 229, 230, 252]. For example, Li *et al*. [183] and Moon *et al*. [229] use relighting techniques to augment InterHand2.6M with more natural lighting and diverse backgrounds.

Hand-object reconstruction: The holistic reconstruction of hands and objects have increased interest in the hand community in recent years [49, 76, 88, 110,

125–127, 171, 200, 330, 335, 377, 412]. Methods in this area mostly assume a given object template and jointly estimate the hand poses and the object rigid poses [76, 110, 125–127, 200, 330, 335, 377, 412] or articulated poses [88]. For example, Cao *et al*. [45] fits object templates to in-the-wild interaction videos. Liu *et al*. [200] introduce a semi-supervised learning framework via pseudo-groundtruth from temporal data to improve hand-object reconstruction.

More recent methods do not assume object templates for hand-object reconstruction [66, 90, 127, 143, 323, 382, 383]. For example, Fan *et al*. [90] introduced the first category-agnostic method that reconstructs an articulated hand and object jointly from a video. Nonetheless,our challenges and insights on hand-object occlusions and camera distortions are still applicable to these more challenging template-free reconstruction settings.

Public reports for the previous challenges (HANDS17 [390] and HANDS19 [11]) distilled the insights from individual review papers and practical techniques into comprehensive summaries to enrich the community's knowledge base. These past challenges use benchmarks which include depth-based hand pose estimation from egocentric views. Instead of depth sensors, the HANDS23 benchmarks are based on affordable and widely applicable image sensors, *i.e*., RGB and monochrome images. This paper further advances the analysis with unique insights, such as multi-view egocentric cameras, object reconstruction in contact, and modeling with recent transformers beyond conventional CNNs.

3.3 HANDS23 Challenge Overview

The workshop contains two hand-related 3D reconstruction challenges in hand-object strongly interacting settings. In this section, we introduce the two challenges and their evaluation criteria.

3.3.1 Workshop challenges

3D hand pose estimation in AssemblyHands: As illustrated in Fig. 3.1, this task focuses on egocentric 3D hand pose estimation from a single-view image based on AssemblyHands. The dataset provides multi-view captured videos of hand-object interaction while assembling and disassembling toy vehicles. In particular,

it provides allocentric (fixed-view) and egocentric recordings and auxiliary cues like action, object, or context information for hand pose estimation. We refer readers to [252] for more dataset details. The training, validation, and testing sets contain 383K, 32K, and 62K monochrome images, respectively, all captured from egocentric cameras. During training, 3D hand keypoint coordinates, hand bounding boxes and camera intrinsic and extrinsic matrices for the four egocentric cameras attached to the headset are provided. The same information is provided during testing, minus the 3D keypoints. Unlike [252], given the availability of multi-view egocentric images, this task lets participants develop multi-view fusion based on the corresponding multi-view images.

Consistent motion reconstruction in ARCTIC: Given an RGB image, the goal of this task is to estimate poses of hands and articulated objects to recover the 3D surfaces of the interaction (see Fig. 3.1). We refer readers to [88] for more details. The ARCTIC dataset contains data of hands dexterously manipulating articulated objects and videos from $8\times$ allocentric views and $1\times$ egocentric views. The official splits of the ARCTIC dataset are used for training, validation, and testing. There are two sub-tasks: allocentric task and egocentric task. In the former, only allocentric images can be used for training and evaluation. For the latter, all images from the training set can be used for training while only the egocentric view images are used during evaluation.

3.3.2 Evaluation criteria

AssemblyHands evaluation: We use MPJPE as an evaluation metric in millimeters, comparing the model predictions against the ground-truth in world coordinates. We provide the intrinsic and extrinsic of the egocentric cameras to construct submission results defined in the world coordinates. Assuming that the human hand has a total N_J joints, we denote wrist-relative coordinates of the prediction and ground-truth as $\hat{J} \in \mathbb{R}^{N_J \times 3}$ and $J \in \mathbb{R}^{N_J \times 3}$, respectively. Given a joint visibility indicator γ_i per joint J_i, we compute the Euclidean distance between predicted and ground-truth joints as $\frac{1}{\sum_{i=1}^{N_J} \gamma_i} \sum_{i=1}^{N_J} \gamma_i \left\| \hat{J}_i - J_i \right\|_2$. The visibility indicator offers per-joint binary labels, representing whether the annotated keypoints are visible from the given egocentric view.

ARCTIC evaluation: Since the original ARCTIC paper [88] has a heavy fo-

cus on the quality of hand-object contact in the reconstructed hand and object meshes, we use Contact Deviation (CDev) introduced in the ARCTIC dataset as the main metric for the competition. In particular, this metric measures the extent to which a hand vertex deviates from the supposed contact object vertex in the prediction. Concretely, suppose that for a given frame, $\{(\mathbf{h}_i, \mathbf{o}_i)\}_{i=1}^{C}$ are C pairs of in-contact ($< 3mm$ distance) hand-object vertices according to ground-truth, and $\{(\hat{\mathbf{h}}_i, \hat{\mathbf{o}}_i)\}_{i=1}^{C}$ are the predictions respectively. The CDev metric is the average distance between $\hat{\mathbf{h}}_i$ and $\hat{\mathbf{o}}_i$ in millimeters, $\frac{1}{C}\sum_{i=1}^{C}||\hat{\mathbf{h}}_i - \hat{\mathbf{o}}_i||$.

For completeness, we report all metrics introduced in ARCTIC. In particular, the task requires the reconstructed meshes to have accurate hand-object contact (CDev) and smooth motion (ACC). Additionally, during articulation or when carrying an object, it is crucial that the vertices of the hand and object which are in contact maintain synchronized movement (MDev). Moreover, we assess hand and object poses, alongside their relative movements, using metrics like MPJPE, AAE, Success Rate, and MRRPE. For detailed information, see [88].

3.4 Methods

This section presents the methods in the two challenges and other competing methods on the leaderboards. Four methods outperform the baseline in both AssemblyHands and ARCTIC.

3.4.1 AssemblyHands methods

Participants develop methods that learn the mapping from egocentric images to 3D keypoints. The methods are categorized into: *heatmap-based* and *regression-based* approaches. Given the presence of complex hand-object interactions in the egocentric scenes, high-capacity transformer models and attention mechanisms addressing occluded regions have been proposed as the backbone networks. Tab. 3.1 summarizes the methods based on the learning, preprocessing, multi-view fusion, and post-processing approaches.

Base: This method uses a *heatmap-based* framework based on heatmaps [230] with 2.5D representations [149] and a ResNet50 [131] backbone. The implementation can be found in [250].

Method	Learning methods	Architecture	Preprocessing	Multi-view fusion (views, phase)
Base	2.5D heatmaps [230]	ResNet50 [131]	-	Simple average (4, test)
JHands [415]	Regression	Hiera [295]	Warp perspective, color jitter, random mask	Adaptive view selection and average (2, test)
PICO-AI	2.5D heatmaps [230] Heatmap voting	RegNety320 [281]	Scale, rotate, flip, tanslate	Adaptive view selection FTL [285] (2, train)
FRDC	Regression 2D heatmaps	HandOccNet [265] with ConvNeXt [204]	Scale, rotate, color jitter	Weighted average (4, test)
Phi-AI	2D heatmaps and 3D location maps [412]	ResNet50 [131]	Scale, rotate, translate, color jitter, gaussian blur	Weighted average (4, test)

Table 3.1: **Method and preprocessing summary in AssemblyHands.** We summarize submitted methods in terms of learning methods, architecture, preprocessing, and multi-view fusion techniques. The tuple (views, phase) indicates the number of views used in either train or test time.

JHands [415]: This method employs a *regression-based* approach with simple MLP heads for regressing 2D keypoints, root-relative 3D keypoints, and the global root depth. The regression training is empowered by a recent fast and strong vision transformer, Hiera [295], pre-trained with masked auto-encoder [129]. A multi-level feature fusion that concatenates the features of different layers is adopted for better feature extraction at different scales. The method additionally uses other publicly available datasets for training, namely FreiHAND [420], DexYCB [49], and CompHand [61]. The implementation is available in [413].

PICO-AI: This method proposes a heatmap voting scheme in addition to the 2.5D heatmaps. Due to their sparsity, the conventional heatmaps pose an imbalance problem between positive and negative samples in the loss function. Hence, the proposed voting mechanism aims to spread the loss evenly across the entire heatmaps. Given the initial guess of keypoints obtained from the heatmaps, the method defines a local region centered on the joint position and operates the soft-argmax within the region to obtain refined keypoint coordinates. This restricts the impact of background points, leading to more reliable optimization. The training is facilitated by CNN-based RegNety320 [281].

Phi-AI: While following the heatmap-based approach, this method adapts MinimalHand [412] with the ResNet50 backbone, where 2D heatmaps and 3D location

maps are regressed. Instead of selecting 3D keypoint coordinates from the location maps, the proposed method modifies it by using heatmap values to weight 3D keypoint coordinates, achieving a more robust estimation. Moreover, the method adds a residual structured layer after the original three-tier cascade networks to refine the calculated location maps. The method further applies the ensemble of final keypoint outputs combined with the Base.

FRDC: This method adopts a hybrid approach by combining *regression* with *heatmap* for training. HandOccNet [265] is modified to regress 3D keypoint coordinates and integrated with an additional branch of 2D heatmap regression. HandOccNet enriches feature extraction with spatial attention mechanisms for occluded regions, making it robust under hand-object occlusions. The method further utilizes a stronger ConvNeXt [204] backbone and feature fusion from the 2D keypoint regressor.

Preprocessing of egocentric images: Compared to conventional static camera setups, egocentric images exhibit unique properties and biases, such as distortion, head camera motion, and different color representations. Thus, it is vital to preprocess egocentric images to alleviate these effects during training. Augmentation techniques are detailed in Tab. 3.1.

The method JHands addresses the distortion issue with a warp perspective operation to make the hands near the edge less stretched. While AssemblyHands provides rectified images converted from fisheye cameras to a pinhole camera model, they often include excessively stretched areas near the edges. To address this, the method calculates a virtual camera and corresponding perspective transformation matrix based on the pixel coordinates of the crop and the camera parameters. The generated crops can be found in the analysis of Fig. 3.3.

Multi-view fusion: Since AssemblyHands offers multi-view egocentric videos, participants can optionally use the constraint of multi-view geometry and fusion techniques during training or inference.

While Base uses a simple average of predicted keypoints from all four camera views in the test time, PICO-AI proposes multi-view feature fusion during training using Feature Transform Layers (FTL) [285]. This FTL training requires fusing two out of four views; thus, the method chooses the most suitable views for every frame. In cases with multiple candidates, the Intersection over Union (IoU) is computed between hand boxes from per-view predictions and 2D keypoints from

previous 3D predictions. The two views with the highest IoUs are selected for their superior prediction reliability.

The methods JHands, FRDC, and Phi-AI apply adaptive fusion in predicted keypoints during testing. The method JHands computes the MPJPE with each other view and selects two results of views with the lowest MPJPE, excluding noisy predictions in the average. If the MPJPE is lower than a threshold, the mean of the two results is calculated as the final result. Otherwise, the result with a lower PA-MPJPE with the predictions in the previous frame is chosen. The methods FRDC and Phi-AI use a weighted average for each view prediction, assigning weights based on each view's validation performance.

Postprocessing: Several postprocessing techniques, including test-time augmentation, smoothing, and model ensembling are used to enhance inference outcomes. In particular, the method JHands applies an offline smooth (Savitzky-Golay) filter on each video sequence.

3.4.2 ARCTIC methods

Tab. 3.2 summarizes the details for each method in terms of the input image dimensions, image backbones, learning rate scheduling, the number of training epochs, batch size, and cropping strategies.

Preliminary: All methods below are regression-based and predict two-hand MANO [293] parameters $\mathbf{\Theta} = \{\theta, \beta\}$ and articulated object parameters $\mathbf{\Omega}$. In particular, with the MANO pose and shape parameters θ, β, the MANO model $\mathcal{H}$ returns a mesh with vertices via $\mathcal{H}(\theta, \beta) \in \mathbb{R}^{778\times3}$. 3D joints are obtained via a linear regressor. The articulated object model $\mathcal{O}$ was introduced in ARCTIC to provide an articulated mesh with vertices via $\mathcal{O}(\mathbf{\Omega}) \in \mathbb{R}^{\mathrm{V}\times3}$, where $\mathbf{\Omega} \in \mathbb{R}^7$ contains the global orientation, global translation, and object articulation.

ArcticNet-SF [88]: Introduced in ARCTIC, it is a single-frame baseline. It first extracts an image feature vector from the input image; then, it regresses hand and object parameters with simple MLPs. The hand and object meshes can then be extracted via $\mathcal{H}(\cdot)$ and $\mathcal{O}(\cdot)$. For more details, see [88].

JointTransformer [2]: JointTransformer enhances ArcticNet-SF by integrating a transformer decoder instead of the MLP regressors for hand and object parameter estimation. The decoder employs learned queries for the angle of each joint, the

Method	Input size	Backbones	Learning rate schedule	Training epochs	Batch size	Cropping
ArcticNet-SF [88]	224 × 224	ResNet50	1e-5	allocentric: 20 egocentric: 50	64	object
DIGIT [91]	224 × 224	HRNet-W32	1e-5	allocentric: 20 egocentric: 50	64	object
AmbiguousHands [271]	224 × 224	ResNet50	1e-5	allocentric: 20 egocentric: 100	32	hand object
UVHand	384 × 384	Swin-L	2e-4 (backbone) 1e-7 (others)	allocentric: N/A egocentric: 50/36*	48	object
JointTransformer [2]	224 × 224	ViT-G	1e-7/1e-5+	allocentric: 20 egocentric: 100	64	object

Table 3.2: **Method and preprocessing summary in ARCTIC.** We summarize baselines on ARCTIC in terms of input dimensions, image backbones, learning rate scheduling, training epochs, batch size and the cropping used for input. *Method trains 50 epochs for decoder and 36 for backbone. +Learning rate is 1e-7 to 1e-4 with linear warmup for first 5% step, and 1e-4 to 1e-7 with cosine decay for rest.

shape and translation of each hand, and the translation, rotation, and articulation of the object. It alternates between self-attention between queries and cross-attention of queries to the elements of the backbone feature map, followed by linear layers that regress the final parameters. Specifically, separate linear layers are dedicated to regressing joint angles, hand shape, hand translation, object translation, object orientation, and object articulation. The best model uses a ViT-G [395] backbone with frozen DINOv2 weights [260].

AmbiguousHands [271]: The method addresses scale ambiguity, resulting from bounding box cropping in data augmentation and camera intrinsics, by employing positional encoding of these elements to mitigate scale issues. This leads to improved spatial alignment. Subsequently, the approach enhances network visibility by integrating local features through distinct hand and object crops. They follow the general approach of ArcticNet-SF to regress hand/object parameters.

UVHand: Since ArcticNet-SF only leverages a global feature vector to estimate hand and object parameters, the image features lack local context. To address this, UVHand leverages Swin-L transformer [203] to extract image features. They then further leverage Deformable DETR [417] to encode the multiple-scale feature maps. The encoded feature maps are then aggregated via self- and cross-attention

before regresssing hand and object parameters.

DIGIT [91]: The method was introduced to estimate strongly interacting hands in [91]. Since ArcticNet-SF is sensitive when the hands are interacting with objects, DIGIT was extended to the ARCTIC setting. Given an image, it first estimates hand part-wise segmentation masks and object masks. The mask predictions are fused with the image features to perform parameter estimation.

Implementation details: Tab. 3.2 shows that all methods use the default cropping as in ARCTIC to crop around the object, while AmbiguousHands performs three crops (around two hands and the object). DIGIT uses the HRNet-W32 backbone [322] and trains with a batch size of 64 with the same learning rate for all iterations. UVHand takes as input a 384×384 image cropped around the object and encodes it with the Swin-L transformer [203] backbone. It was trained with a batch size of 48 with a learning rate of 2e-4 for the backbone and 1e-7 for other weights. Due to computational cost, they train 50 epochs for the decoder and 36 for the backbone. JointTransformer uses ViT-G [395] backbone with frozen DINOv2 [260] weights to train with a batch size of 64. It performs a linear warmup from 1e-7 to 1e-4 in the first 5% steps and uses cosine decay from 1e-4 to 1e-7 for the rest of the steps.

3.5 Results and Analysis

3.5.1 Results

Here, we benchmark results of valid submissions for state-of-the-art comparison in AssemblyHands and ARCTIC and other more recent baselines. In particular, for AssemblyHands, we report egocentric hand pose estimation results. For ARCTIC, we report results for the allocentric and egocentric test sets.

AssemblyHands benchmark: Tab. 3.3 shows the final test scores on the AssemblyHands dataset. The methods in the table exceed the baseline (Base) with a test score of 20.69 MPJPE. Notably, the methods JHands and PICO-AI achieve a nearly 40 % reduction over the baseline. The methods FRDC and Phi-AI improve the test score by 20.3 % and 16.5 % against the baseline, respectively.

ARCTIC benchmark: Tab. 3.4 presents the comparative performance of methods in the ARCTIC dataset, where ArcticNet-SF serves as the initial benchmark.

Method	Score	Hand distance (px)			Verb class						
		-200	200-250	250-	clap	inspect	pass	pick up	position	position screw on	pull
Base	20.69	20.31	21.97	24.85	19.88	22.89	22.48	21.85	22.65	21.62	18.74
JHands	12.21	12.35	11.98	13.72	10.65	16.27	12.86	13.67	14.58	12.8	11.06
PICO-AI	12.46	12.51	11.62	12.95	12.98	15.3	11.37	13.2	13.18	11.39	15.13
FRDC	16.48	16.39	15.89	18.69	15.24	21.33	17.86	18.26	19.21	18.03	12.83
Phi-AI	17.26	17.24	15.81	19.51	19.86	20.93	17.91	19.01	19.7	19.35	17.3

	Verb class (continue)									
	push	put down	remove	remove screw from	rotate screw	tilt down	tilt up	unscrew	none	
Base	19.29	20.26	19.99	16.47	22.71	22.95	13.12	15.11	20.78	19.82
JHands	13.96	13.72	13.11	11.72	12.26	14.11	9.61	9.92	12.25	10.99
PICO-AI	12.29	13.41	12.83	9.99	13.7	13.72	10.44	11.56	12.87	11.71
FRDC	19.52	17.87	18.55	14.47	16.44	18.81	14.03	13.37	16.41	15.01
Phi-AI	19.12	18.29	18.18	13.95	18.35	19.78	13.13	15.36	17.29	15.8

Table 3.3: **Method performance in AssemblyHands.** We compare AssemblyHands method performance on egocentric test data. We show the final MPJPE on the test set as the metrics (lower better). We also provide detailed evaluations, regarding the varying distances of hand position from the image center and different verb action categories. The hand distance is computed by the distance from the image center to the hand center position per image, and averaged over the lower two views of the headset. Verb classes of "attempt to X" are merged to "X" for simplicity. The higher and lower three verbs are color-coded in red and blue, respectively.

		Contact and Relative Positions		Motion		Hand	Object	
	Method	CDev $[mm]\downarrow$	MRRPE$_{rl/ro}$ $[mm]\downarrow$	MDev $[mm]\downarrow$	ACC$_{h/o}$ $[m/s^2]\downarrow$	MPJPE $[mm]\downarrow$	AAE $[^\circ]\downarrow$	Success Rate $[\%]\uparrow$
Allocentric	ArcticNet-SF	41.56	52.39/37.47	10.40	5.72/7.57	21.45	5.37	71.39
	DIGIT	34.92	44.19/35.43	**8.37**	**4.86**/6.63	17.92	5.24	76.52
	UVHand	64.15	84.68/70.31	14.12	7.05/12.04	40.99	12.36	31.47
	AmbiguousHands	33.25	45.78/34.56	10.12	6.37/6.40	18.02	4.64	81.94
	JointTransformer	**27.97**	**36.17/28.18**	8.93	6.08/**5.79**	**17.12**	**3.95**	**89.79**
Egocentric	ArcticNet-SF	44.71	28.31/36.16	11.80	5.03/9.15	19.18	6.39	53.89
	DIGIT	41.31	25.49/32.61	**9.48**	4.01/8.32	16.74	6.60	53.33
	UVHand	40.43	40.93/36.88	9.96	5.32/8.33	24.53	7.32	57.28
	AmbiguousHands	35.93	**23.07**/27.53	9.51	**3.95/6.76**	**16.26**	4.86	68.36
	JointTransformer	**32.56**	26.07/**26.22**	11.34	5.52/8.68	16.33	**4.44**	**74.07**

Table 3.4: **Method performance in ARCTIC**. We compare performance in both allocentric (top half) and egocentric (bottom half) views. We evaluate using metrics for contact and relative position (measuring hand-object contact and prediction of relative root position), motion (assessing temporally-consistent contact and smoothness), and hand and object metrics (indicating root-relative reconstruction error). We use the CDev score as the main metric for this competition. We denote left and right hands as l and r, and the object as o.

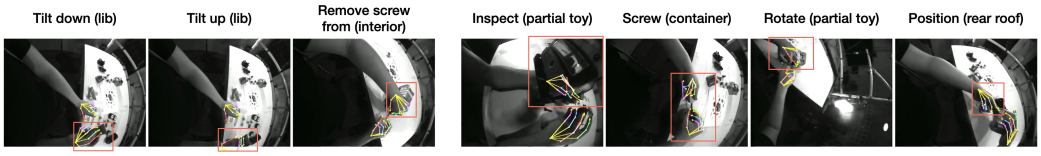


Figure 3.2: **Qualitative results per action in AssemblyHands**. We show Base results with "verb (noun)" actions. The left three figures are lower error situations while the right four ones are failure cases. The red boxes denote the area where the action occurs.

The majority surpass ArcticNet-SF in both allocentric and egocentric views, except for UVHand, which underperforms due to incomplete training. In the egocentric view, AmbiguousHands excels in creating smooth, consistent mesh motions (refer to MDev and $\text{ACC}_{h/o}$ metrics). Notably, JointTransformer stands out by significantly lowering CDev errors by 32.7% in allocentric and 27.2% in egocentric settings compared to the baseline.

3.5.2 AssemblyHands analysis

We provide analysis, regarding action-wise evaluation, distortion effect in training, and the effect of multi-view fusion.

Action-wise evaluation: To analyze errors related to hand-object occlusions and interactions, we show pose evaluation according to fine verb action classes in Tab. 3.3. We use the verb classes annotated by Assembly101 [303], spanning every few seconds in a video. Fig. 3.2 shows qualitative results of representative verb classes with the top and bottom error cases.

We observe that the performance varies among different verb actions. The verbs "tilt down/up" and "remove screw from" exhibit lower errors among the submitted methods, because hands are less occluded and their movement is relatively stable. The "tilt" action holds a small part of the toy and turns it around alternately, leading to less overlap between the hand and the object (lib). The "remove screw from" action takes a screw out from the toy vehicle by their hand where observed hand poses do not change drastically.

Higher error classes, such as "inspect", "screw", "rotate", and "position", contain heavy occlusions, fast hand motion, complex two hands and object interac-

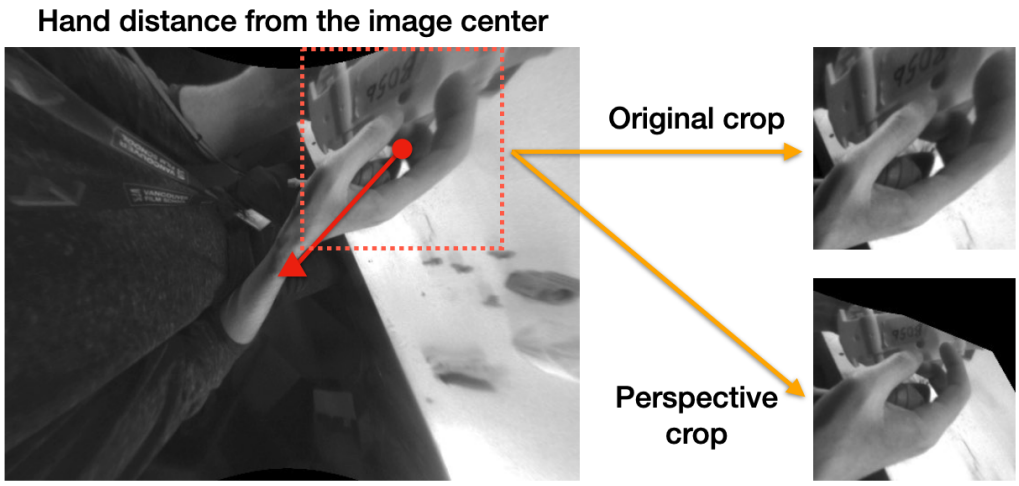


Figure 3.3: **Effect of distortion in AssemblyHands**. The officially released images in the dataset have highly stretched areas near the edges (original crop). The method JHands uses a perspective crop with a virtual camera to correct this distortion.

tions. The "inspect" action brings the toy close to the human eyes where the toy occupies a large portion of the image causing heavy object occlusions. The "screw" action involves intricate fingertip movements to rotate the screwdriver quickly. The "rotate" and "position" actions are performed so that the two hands and the object interact in close contact, which complicates the estimation. We observe that the top two methods JHands and PICO-AI significantly correct the results of these higher error actions compared to the other submitted methods.

Bias of hand position in an image: Hands near image edges are highly distorted due to the fish-eye cameras. Directly using these noisy images in training will degrade performance [414]; thus, some methods create new crops with less distortion, select training instances, or adaptively fuse predictions during the inference. Specifically, Fig. 3.3 shows that the method JHands reformulates the perspective during cropping and creates less-distorted (perspective) crops.

To study this effect in the final performance, we split the evaluation instances into classes with different 2D distances between the hand center and the image center in Tab. 3.3. Higher distances (250- pixels) indicate closer hand crops to image edges. The method, Base, without any training instance selection and dis-

View	MPJPE	Miss(%)
cam1	37.97	70.8
cam2	25.71	88.3
cam3	22.19	0.92
cam4	22.29	0.74
cam3+4	21.52	0.08
all four	20.69	0

Table 3.5: **Multi-view fusion in AssemblyHands.** We use the Base result to show performance before and after fusion. Missing instances per view are denoted as "Miss(%)".

tortion correction, has higher error as the crops approach image edges (20.31 → 24.85). In contrast, the newly proposed methods are more robust and have a lower error, particularly in the 200-250 range. We observe that the ranges 200-250 and 250- occupy 10% and 5% of the test images, respectively, thus the improvement in the 200-250 range helps the lowering of the overall score.

Effect of multi-view fusion: The multi-view egocentric camera setup is unique to the dataset. We show the statistics and performance of multi-view fusion in Tab. 3.5. Note that Tab. 3.3 shows the final results after multi-view fusion.

We found that samples captured from the lower cameras (cam3 and cam4; see Fig. 3.1 for the layout) are numerous (fewer missing samples) and their errors are lower as they are faced toward the area occurring hand interactions. Conversely, the samples from cam1 and cam2 are fewer and unbalanced as their cameras often fail to capture hands due to the camera layout. For instance, cam1 (top-left) tends to capture more hand region than cam2 as the participants are mostly right-handed and bring up the object with the right hand, which can be better observed from cam1. Given this uneven sample distribution, the proposed adaptive view selection methods in either training or testing are essential to perform effective multi-view fusion, and outperform the Base's test-time average using all views all the time (see **Multi-view fusion** in Sec. 3.4.1).

We further study the performance gain before and after multi-view fusion using Base's results. While per-view performance achieves 22.19 and 22.29 in cam3 and cam4, respectively, their fused results with simple average reduce the error to

21.52. Merging all four views has shown to be more effective than two-view fusion (20.69 *vs*. 21.52), indicating a 6.5% reduction compared to the single camera setup (cam3). This suggests predictions from the top views (cam1 and cam2) are informative in averaging even when they are prone to be erroneous.

Qualitative results: Fig. 3.4 shows the qualitative results of submitted methods and failure patterns indicated by the red circles. The left hand in the first row grabs the object where the left thumb finger is only visible. While Base fails to infer the plausible pose, JHands enables estimation in such heavy hand-object occlusions compared to the GT. However, the methods PICO-AI and FRDC incorrectly predict the location of the left thumb finger and Phi-AI's prediction of the left index and middle fingers is also erroneous. The second row is the case where two hands and an object are closely interacting, particularly the left thumb finger presents near the right hand. The methods Base, FRDC, and Phi-AI fail to localize the left thumb finger. The third and fourth rows indicate hand images presented near the image edges. The methods Base, PICO-AI, and Phi-AI are prone to produce implausible predictions, including noise and stretched poses due to the distortion effect. The method JHands with distortion correction successfully addresses these edge images.

Per-view analysis: Fig. 3.5 shows the detailed statistics and performance of per-view predictions. Considering per-sequence results, we find the sample availability (blue bars) and performance (green bars) from cam1 and cam2 vary among different users. In contrast, the number of samples and performance of cam3 and cam4 are mostly stable. This study further necessitates the sample selection and multi-view fusion adaptively for each sequence (user).

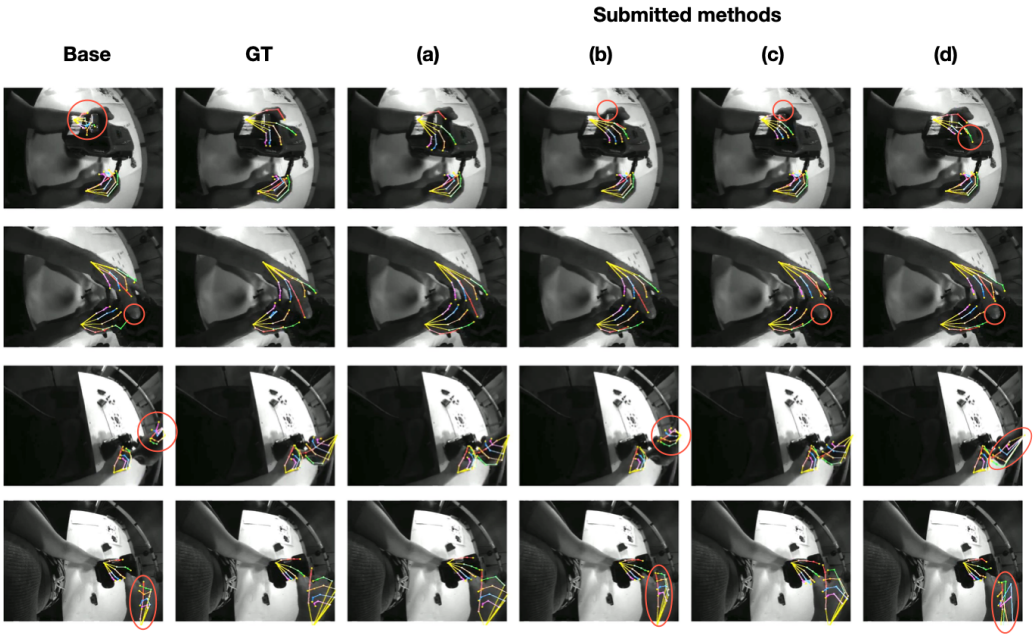


Figure 3.4: **Qualitative results of submitted methods in AssemblyHands.** The columns correspond to the results of Base, ground-truth (GT), submitted methods, namely (a) JHands, (b) PICO-AI, (c) FRDC, and (d) Phi-AI. The red circles indicate where failures occur.

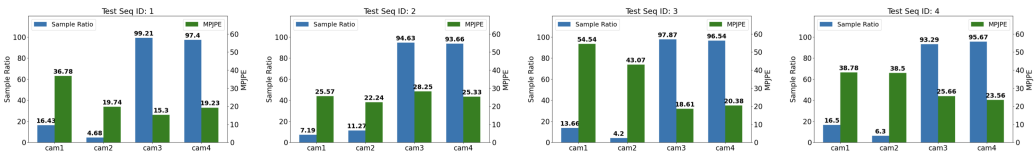


Figure 3.5: **Results of multi-view fusion in AssemblyHands**. We analyze the availability of samples and performance per camera view. The two lowest cameras (cam3, cam4) out of the four cameras allow us to capture hands most of the time (>93 % of samples). In contrast, the images from cam1 and cam2 are fewer and the error varies in different sequences.

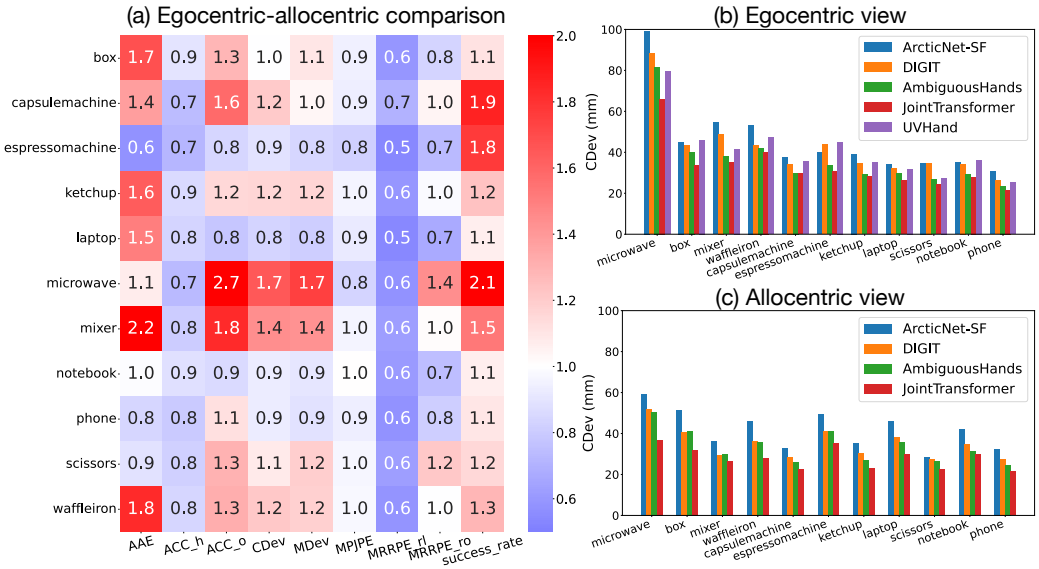


Figure 3.6: **Performance comparison: Egocentric vs Allocentric**. (a) Comparative difficulty ratio of egocentric to allocentric views. (b) Egocentric view performance by method across objects. (c) Allocentric view performance by method across objects.

3.5.3 ARCTIC analysis

Egocentric-allocentric comparison: Fig. 3.6a compares the performance between egocentric and allocentric views. In particular, we compute a ratio between the metric values of the egocentric and allocentric view to measure the extent of difficulty for the egocentric view compared to the allocentric view. Since success rate is a metric whose value is positively correlated to performance, we take its reciprocal ratio. We average the ratios across methods and actions.

We observe that hand pose-related metrics such as MPJPE and ACC_h are less than 1.0 on average (see blue color cells), meaning the the egocentric view is easier than allocentric view. This is because most allocentric cameras in ARCTIC are meters away from the subject while the egocentric camera is often close-up, offering higher hand visibility. Relative translation metrics between hand and object such as MRRPE_{rl} are also easier in the egocentric view because estimating translation is more difficult from further cameras.

Object reconstruction performance faces unique challenges, as highlighted by

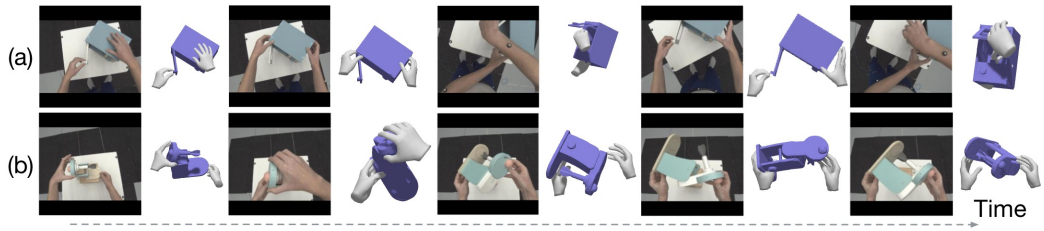


Figure 3.7: **Egocentric reconstruction by top method in ARCTIC**. In the egocentric view, object reconstruction struggle when the object is partially observed on the image boundaries, as well as when heavy hand/arm occlusion occurs.

the red cells. In the egocentric view, it is notably more difficult to reconstruct accurate object surfaces, articulation (AAE), and hand-object contact (CDev). This increased difficulty arises because objects are often positioned at the image edges and obscured by human arms. Additionally, object poses exhibit greater diversity in the egocentric view due to varying camera angles and occlusions. While a static camera maintains consistent camera extrinsics across a sequence, an egocentric camera's extrinsics change with each frame, resulting in higher diversity in camera-view object 6D poses. This diversity complicates object pose estimation in egocentric views. Fig. 3.7 illustrates these challenges using the best-performing method, JointTransformer. Despite achieving reasonable hand poses, object poses are significantly impacted by occlusions from hands and arms, and the egocentric view undergoes substantial changes throughout a sequence.

Object-wise evaluation: Fig. 3.6b and 3.6c break down performances on different objects in the egocentric and allocentric view test sets. The best method in both cases is JointTransformer. The hardest object to reconstruct with good contact consistency (see CDev) is the microwave in both settings due to global rotation sensitivity (see Fig. 3.7), though this can be mitigated by a keypoint-based approach [117]. Estimating objects in the egocentric view is more difficult than in the allocentric view, which is indicated by higher errors for all methods.

Action-wise evaluation: Fig. 3.8 compares performance of different methods in "grab" and "use" actions. In ARCTIC, there are sequences to interact with the object with two types of actions by either not articulating the object, or allowing object articulation. Interestingly, the "grab" motion is more challenging in egocentric and allocentric views. We hypothesize that this is because there are

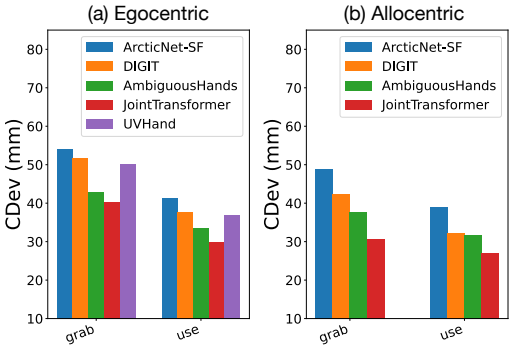


Figure 3.8: **Hand-object contact quality for reconstructed results per action**. We evaluate the contact quality of the 3D reconstruction results from all methods for each action (*i.e.*, grab or use), using Contact Deviation (CDev) in mm as the metric, where lower values indicate better quality.

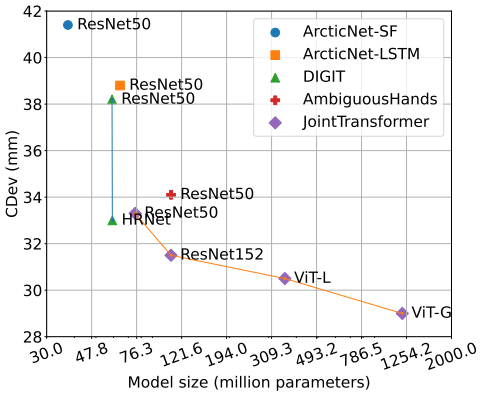


Figure 3.9: **Contact deviation vs. model size**. We assess the contact quality of the reconstruction results, varying by the number of parameters in each model. Contact quality is measured using Contact Deviation (CDev) in mm, with lower values indicating superior results.

more diverse object poses for "grab" motions since during object articulation, the participants often focus on articulation instead of object manipulation.

Effect of model size: Fig. 3.9 illustrates the impact of model size on hand-object contact performance, measured by CDev, for reconstruction on the allocentric validation set. Most methods utilize ResNet50, with JointTransformer being the top performer. As the trainable parameters in the backbone increase, JointTransformer consistently reduces the CDev error. Note that the x-axis is in log-scale. JointTransformer achieved a CDev error of 30.5mm with ViT-L and 29.0mm with ViT-G, which has ten times more parameters than ViT-L. Interestingly, JointTransformer uses frozen weights in the large-scale ViT-L and ViT-G backbones, yet achieves the best results. This suggests a potential direction for leveraging large-scale foundational backbones for hand-object reconstruction.

3.6 Conclusion

In this paper, we introduce the HANDS23 challenge and provide analysis based on the results of the top submitted methods and more recent baselines on the leaderboards. We organize and compare the submissions and their implementation details based on the learning methods, architecture, pre- and post-processing techniques, and training configurations. We thoroughly analyze various aspects, such as hand-object occlusions, action and object-wise evaluation, distortion correction, multi-view fusion, egocentric-allocentric comparison, and performance gain of large transformer models.

Future directions: There are several future directions that the community can take. For example, one can explore more efficient training using multi-view egocentric cameras, leveraging 3D foundation priors [199, 269] to regularize template-free hand-object reconstruction [90], estimating hand-object poses with more expressive representations (*e.g.*, heatmap-based approaches [117]), incorporating motion and temporal modeling [98], featuring more diverse egocentric interaction scenarios, recognizing actions through captured hand poses [304], learning robotic grasping from reconstructed hand-object pose sequences, and so forth.

Chapter 4

Hand Self-contact Benchmark and Generative Pose Modeling

One can hardly model self-contact of human poses without considering underlying body shapes. For example, the pose of rubbing a belly for a person with a low BMI leads to penetration of the hand into the belly for a person with a high BMI. Despite its relevance, existing self-contact datasets lack the variety of self-contact poses and precise body shapes, limiting conclusive analysis between self-contact poses and shapes. To address this, we begin this chapter by introducing the first extensive self-contact dataset with precise body shape registration, **Goliath-SC**, consisting of 383K self-contact poses across 130 subjects. Using this dataset, we propose generative modeling of self-contact prior conditioned by body shape parameters, based on a body-part-wise latent diffusion with self-attention. We further incorporate this prior into single-view human pose estimation while refining estimated poses to be in contact. Our experiments suggest that shape conditioning is vital to the successful modeling of self-contact pose distribution, hence improving single-view pose estimation in self-contact.

4.1 Introduction

Human poses in our daily life often involve *self-contact*, such as face touching, arm crossing, or hand placement, where body parts come into contact with the body surface. These interactions with our own body are not only unconsciously

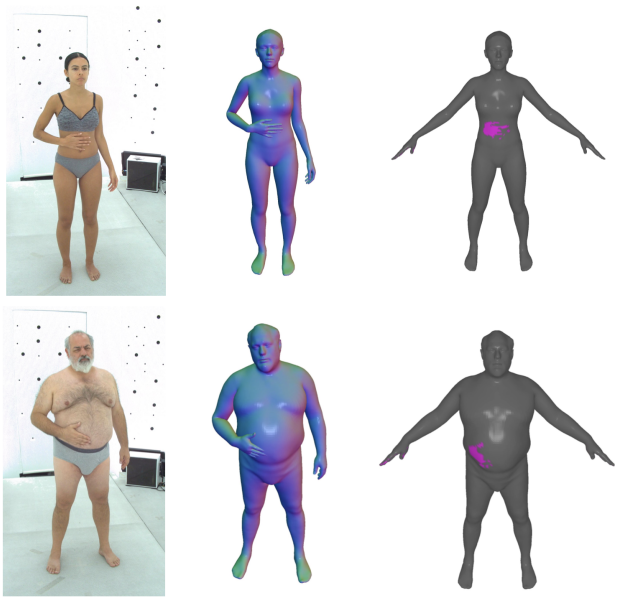


Figure 4.1: **Body shape dependency in self-contact poses**. We observe that self-contact poses (*e.g.*, "rubbing belly") are influenced by the subject's body shape; for example, a person with a slimmer body (top) engages in different self-contact poses over one with a larger torso (bottom). Indeed, the contact maps on the template mesh (right) are presented differently. Examples are sampled from the **Goliath-SC** dataset we captured.

displayed but also carry profound meaning across various disciplines, including psychology and social communication. Observing self-contact gestures, the areas touched can signal emotional states (*e.g.*, anxiety or tension) [121, 122, 267], while expressing linguistic symbols and contexts in sign language [73, 342, 387]. Notably, these self-contact poses are inherently constrained by the underlying body shapes. As shown in Fig. 2.1, two subjects performing "rubbing belly" gestures exhibit different poses and contacts due to variations in body shapes and proportions related to Body Mass Index (BMI). Despite their significance, accurately modeling self-contact poses remains a considerable challenge [94, 236]; particularly, the dependency of human poses on body shapes is underexplored.

The challenge of self-contact modeling stems from the lack of datasets containing large self-contact poses with precise body shape registration. Existing 3D body self-contact datasets, HumanSC3D [94] and MTP [236], contain small

self-contact poses (1-4K poses) and suffer from inaccurate registration due to the lack of paired RGB images [236]. Other studies have highlighted specific part interactions, such as hand-hand [230] or hand-face contact [307]. However, their scopes are limited to isolated body parts and fail to capture the holistic nature of self-contact, overlooking how the full-body pose and shape influence the contact.

Given the limitations of the existing datasets, we begin by offering the first extensive self-contact dataset with varying full-body poses and precise body shape registration, dubbed **Goliath-SC**. Our self-contact dataset contains the largest amount of self-contact poses, comprising 383K poses from 130 subjects. Additionally, it provides accurate full-body mesh registration based on 3D scans in a multi-camera dome (Goliath [215]), which are converted to SMPL-X [266] to access body shape parameters.

Using this dataset, we model the dependency of self-contact poses on body shapes via generative models. The generative modeling is designed to learn self-contact pose distribution for the given body shapes, independent of image input. Pose-based generative training has an advantage over direct pose regression from images [94, 230, 236, 307] due to its generalizability to unknown environments and subjects. Removing image input helps debias from the captured environments. Furthermore, it facilitates interpolation in the learned pose–shape space, enabling the modeling of plausible self-contact poses for novel body shapes or contact locations not explicitly seen during training.

In more detail, our approach involves a new insight of *shape-dependency* in generative modeling with denoising diffusion. Unlike joint distribution modeling between pose and shape [173, 237] of 3D human models [206, 293], we explicitly model the shape-dependent manifold of self-contact poses using diffusion models [137, 314]. Specifically, we develop a latent diffusion model with self-attention, termed **PAPoseDiff**, which considers the relationship among highly interacting body parts (*e.g.*, hands, body, and face).

Finally, we leverage the learned diffusion prior to refine 3D poses in self-contact. Given the initial SMPL-X estimation, we refine the poses to have a smaller error to the 2D keypoint observation, while maintaining the plausibility in contact acquired by the former generative training. Our experiments demonstrate that our refinement with the shape-conditional diffusion prior surpasses a recent diffusion prior for human contact (BUDDI [237]) and the state-of-the-art

foundation model with direct regression (SMPLer-X [41]) in the Goliath-SC *eval* set with unseen subjects.

Our contributions are summarized as follows:

- We introduce a new self-contact dataset **Goliath-SC** with extensive poses and precise body shape registration.
- We propose generative learning of the shape-dependent manifold of self-contact poses, along with a latent diffusion with part-aware self-attention, **PAPoseDiff**.
- We propose an efficient single-view pose refinement, fitting initial SMPL-X predictions to the observed 2D keypoints using the learned diffusion model.

4.2 Related Work

Self-contact datasets: Human contact is taken into account in 3D human reconstruction [160, 236, 385]. These studies include self-contact of a single person (*e.g.* crossing arms) [94, 236], multi-person interactions like hugging [93, 136, 160, 237, 385], or contact with external environments, such as scene [123, 141, 189] and handheld objects [64, 88, 110, 194, 198, 251, 252, 324]. However, constructing self-contact datasets is particularly difficult due to higher self-occlusion. HumanSC3D [94] and MTP [236] datasets contain a limited number of poses (1-4K poses) and inaccurate annotations due to the absence of paired RGB images with the captured poses [236]. Other studies focus on specific part interactions; InterHand2.6M captures hand-hand interactions [230], while Decaf highlights hand-face contact [307]. Despite allowing fine interaction analysis, capturing only isolated parts disregards the holistic perspective of self-contact, *i.e.*, how the body influences hands and face in contact. In contrast, our captured Goliath-SC dataset provides extensive self-contact poses (383K) with a dense camera setup, including full-body registration with precise shapes. Our dataset also includes continuous pose variations, unlike frame-independent pose registration in MTP [236].

Self-contact estimation: Previous self-contact works follow a *regressive* approach, aiming to estimate contact states from a single image. Early attempts [94, 236] rely on the annotation of discrete 2D contact labels, representing which body parts are in contact, though the annotation process is labor-intensive and difficult

to scale. Fieraru *et al*. formulate the tasks of segmenting in-contact parts and predicting interacting part pairs (contact signature) [94]. Muller *et al*. estimate human poses in self-contact [236], with two distinct training configurations: (1) supervised training of a regressor on 3D GTs (*i.e*., MTP [236]) and (2) additional optimization when 3D GTs are unavailable, relying instead on discrete contact labels (*i.e*., in-the-wild data like DSC [236]). Without using such contact labels, recent human foundation models (*e.g*., SMPLer-X [41]) extend (1)'s approach to train a ViT network [83] across various 3D human datasets, including self-contact scenarios (*i.e*., HumanSC3D [94] and MTP [236]). While this simple regression strategy generalizes across domains, it still struggles to capture the nuanced self-contact. To address this, we investigate a *generative* approach to refine the regressor's estimates, without relying on image input and manual annotations for 2D contact parts.

Diffusion models: Denoising diffusion [137, 314] is becoming a popular choice for generative prior modeling due to its higher capability compared to handcrafted methods [3, 266] or VAEs [164, 266]. Diffusion models are trained to iteratively denoise a Gaussian noise to sample from the learned data distribution [137, 314]. While they have been widely adopted, *e.g*., for motion synthesis [331, 370], only a few works have modeled human contact using diffusion models. BUDDI [237] and InterHandGen [173] are concurrently proposed to model the contact between two bodies (either human bodies or hands) with the DDPM formulation [137].

While these previous methods learn the joint distribution of the pose and shape parameters of SMPL [206] or MANO [293], our diffusion modeling relies on a new assumption that *pose should depend on body shape*, thus generating poses conditioned on the given body shapes. Furthermore, when adapting the diffusion prior to single-view pose estimation, our proposed refinement-based method does not require additional fine-tuning as in [173].

4.3 Self-Contact Analysis and Dataset

We introduce a new self-contact dataset with varying full-body poses and shapes, termed **Goliath-SC**. Our capture is based on a multi-camera dome setup of Goliath [215] with 3D full-body scans from 220 RGB cameras. The scope of captured activities lies in natural self-contacts occurring in daily life like touching the

Figure 4.2: **Examples of our Goliath-SC dataset and contact heatmap.** We capture self-contact poses from 130 subjects with scripted action instructions (*e.g.*, “hand hitting forehead”). Examples are sampled from the subjects of Goliath-4 [215].

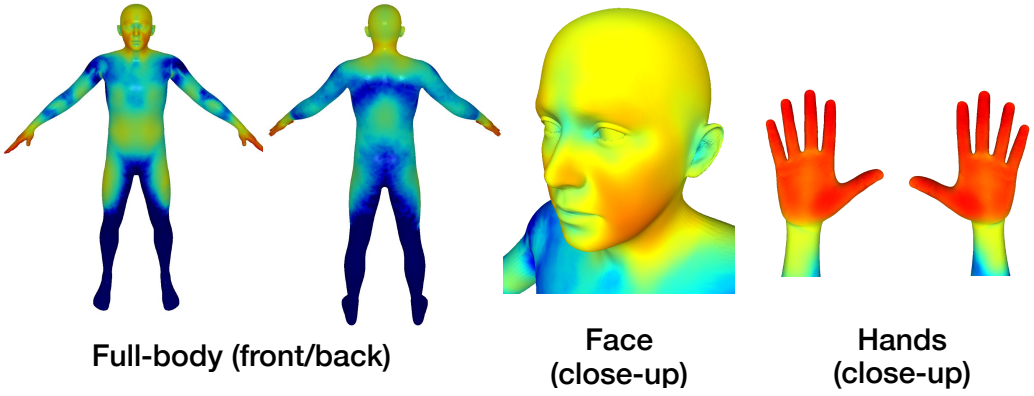


Figure 4.3: **Contact heatmap of our Goliath-SC dataset.** We compute vertex-wise binary contact maps to find contact frames, and the averaged heatmap is shown.

Dataset	#SCPose	#Subj.	Params	Annot.
HumanSC3D [94]	1.0K	6 (3/3/0)	SMPL-X [266] / GHUM [368]	Mocap
FlickrSC3D [94]	<3.9K	-	SMPL-X	Pseudo-3D-GTs
MTP [236]	1.6K	148 (52/96/0)	SMPL-X	Pseudo-3D-GTs
Goliath-SC (Ours)	383K	130 (70/56/4)	SMPL-X	MV RGB scan [215] (cam: 220)

Table 4.1: **Comparison of full-body self-contact datasets.** We compare the number of self-contact poses, captured subjects, body parametrization, and annotation methods. The subject data include the gender ratio (female/male/non-binary).

face, body, hands, etc. Tab. 4.1 and Fig. 4.2 show data statistics and examples.

Our dataset has the following advantages. (1) Our captures contain substantial self-contact data with 383K poses from 130 subjects, exceeding the existing self-contact datasets [94, 236] by two orders of magnitude. (2) Owing to numerous cameras with increased resolutions for the face and hands areas [215], it provides high-quality mesh registration with fine details for hands and face. This enables capturing fine self-contacts like "rubbing eyes" and "massaging hands", which is distinguished from the existing self-contact scenarios [93, 94, 236, 385]. (3) Instead of collecting in-contact poses independently [236], we capture the sequence of natural self-contact poses at 30 Hz with scripted action instructions (*e.g.*, "rubbing belly"), leading to diverse and continuous self-contact poses. (4) Unlike targeting specific body parts (*e.g.*, hand-hand in InterHand2.6M [230], hand-face in Decaf [307]), our dataset provides complete 3D full-body poses including hand-hand and hand-face interactions. This enables holistic behavior modeling in which hands and face are constrained by the body's kinematics. (5) To model the shape-dependent manifold, we convert the registered meshes to SMPL-X [266], which gives the latent shape parameters.

Contact maps and data screening: To comprehend self-contact patterns, we compute vertex-wise contact maps; see Fig. 4.3. We first create binary contact maps by discriminating if the hand vertices are close ($<$ 3mm) to the rest of the

body vertices, and then select contact frames with positive contact maps of the captured sequence. This indicates that each sample corresponds to a unique self-contact pose in which the hand is touching somewhere on the body.

We then calculate the contact heatmap from the binary maps, indicating contact likelihood as [324]. We observe that it includes various interactions across hands, face, neck, arms, and torso. While the studies on hand-object grasping have a high contact likelihood in the finger areas [88, 89, 252, 324], our self-contact data include frequent interactions in the palm of hands as well. This suggests that self-touching gestures are expressed by using hands widely from the palm to the fingertips.

4.4 Method

We present our proposed generative diffusion model for self-contact pose modeling. Here, we aim to model the manifold of self-contact poses, particularly depending on the subject's body shape. We detail our task and model setup in Secs. 4.4.1 and 4.4.2 and training objectives in Sec. 4.4.3. We then provide the inference process in Sec. 4.4.4. Figs. 4.4 and 4.5 show the overview of our proposed diffusion model, dubbed **PAPoseDiff**, and our refinement scheme to obtain refined pose $\boldsymbol{X}_0^{ref}$ given initial 3D pose estimate $\boldsymbol{X}_0^{init}$, respectively.

4.4.1 Diffusion process

We follow the DDPM formulation [137], where the diffusion process consists of forward and reverse paths spanned with diffusion time steps $t \in [1, T]$. The forward process ($1 \rightarrow T$) takes an input data $\boldsymbol{X}_0$ and gradually adds standard Gaussian noise ϵ_t to the data. We denote the process of diffusing $\boldsymbol{X}_0$ at step t as $\boldsymbol{X}_t = \text{noise}\left(\boldsymbol{X}_0, t\right)$, formulated as

$$\text{noise}\left(\boldsymbol{X}_0, t\right) = \sqrt{\bar{\alpha}_t}\boldsymbol{X}_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon_t. \tag{4.1}$$

The noisiness of $\boldsymbol{X}_t$ is controlled by noise variances β_t, *e.g.*, given by a cosine scheduler [245]. The coefficients that balance the noise and data terms are determined by $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \Pi_{s=1}^{t}\alpha_s$. T is set to 1000.

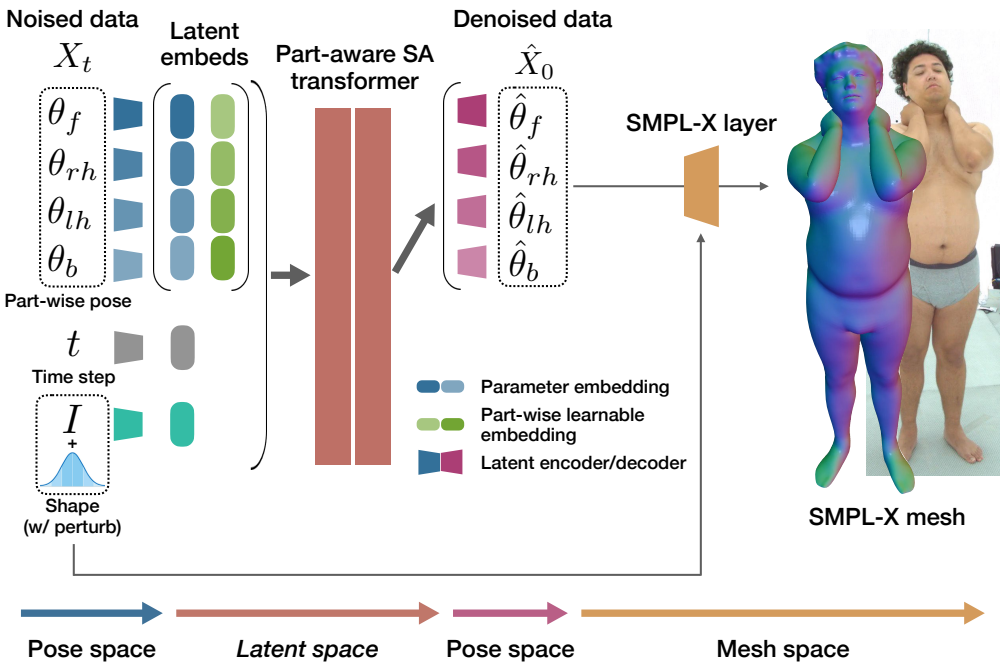


Figure 4.4: **Shape-conditional denoising diffusion model for self-contact poses.** Our proposed diffusion model, **PAPoseDiff**, follows latent diffusion with part-aware attention. The model is trained to generate part-wise pose parameters conditioned on the shape information while considering their interactions with self-attention (SA). We also add small perturbations for the shapes to generalize to unseen subjects. The training losses are described in Sec. 4.4.3.

The reverse process ($T \rightarrow 1$) denoises the data in every step and finally generates a clean sample $\hat{\boldsymbol{X}}_0$ with a learnable model f. Following [314, 331], given the noised data $\boldsymbol{X}_t$ in step t, the model directly approximates the original data $\boldsymbol{X}_0$ as $\hat{\boldsymbol{X}}_0 = f(\boldsymbol{X}_t, t)$. For conditional generation, the model can take an additional conditional input $\boldsymbol{c}$ as $f(\boldsymbol{X}_t, t, \boldsymbol{c})$.

4.4.2 Shape-dependent pose modeling

With the diffusion formulation, we construct data representation for poses and network architectures that enforce the shape-dependent constraint. We explicitly model the interactions between different body parts including hands, body, and face, which is a distinction from learning body parameters only in the previous

studies of human contact [236, 237].

Data representation: To learn self-contact poses, we use part-wise pose parameters as the target to denoise and shape parameters as conditional input, which are obtained by the differentiable SMPL-X model [266]. The input pose data are constructed as $\boldsymbol{X} = [\boldsymbol{\theta}_f, \boldsymbol{\theta}_{rh}, \boldsymbol{\theta}_{lh}, \boldsymbol{\theta}_b]$, where $\boldsymbol{\theta}_f \in \mathbb{R}^{3+10}, \boldsymbol{\theta}_{rh}, \boldsymbol{\theta}_{lh} \in \mathbb{R}^{15\times3}, \boldsymbol{\theta}_b \in \mathbb{R}^{21\times3}$ indicate pose parameters for face, right hand, left hand, and body, respectively. To clarify, the target face parameters, jaw pose and expression, are combined in $\boldsymbol{\theta}_f$ for convenience.

The shape parameters are represented as $\boldsymbol{I} \in \mathbb{R}^{N_s}$ of the SMPL-X where N_s is the shape dimension ($\leqslant 300$). These encode the subject's physical identity such as bone length and body size. In contrast to part-isolated input such as hand-hand [230] (MANO's) or hand-face interactions (MANO's + FLAME) [307], the whole body parametrization provides additional constraints about the location of hands and face restricted through the kinematic chain of the body. Owing to this, our representation only relies on local pose for simplicity while global orientation and translation of SMPL-X are disregarded.

Latent diffusion with part-aware self-attention: Regarding part-wise interaction modeling, we propose a part-aware self-attention transformer with latent embedding as f. Unlike learning on pose parameters directly in the diffusion process [173, 237], we train the diffusion model in the latent space, inspired by latent diffusion for image synthesis [291]. As joint movements in human motion are highly coordinated, the DOFs of whole-body joints can be represented in a lower dimensional space [181, 186, 293]. Similarly, self-contact poses are intrinsically embedded in a latent manifold, as restricted to move along the body surface. To enforce these constraints, we utilize auto-encoders for part-wise poses, enabling the discovery of plausible and semantically meaningful latent embeddings in training.

With the embeddings of the pose, diffusion time, and shape, we then utilize a self-attention transformer [344] as the denoising module. Specifically, the query, key, and value of the attention are given by the concatenation of the embeddings across face, right/left hand, body, time, and shape. This enables considering the interactions across part-wise poses, shape, and diffusion time (*i.e.*, degree of given noisiness). Part-wise learnable embeddings are also added to facilitate part-aware relational learning as [237].

Shape-conditional perturbation: Instead of naive conditioning of shape parameters, enriching the diversity of subjects in training is critical in learning the shape-dependent manifold and generalizing to unseen subjects in testing. Prior works use conditional dropout $\boldsymbol{c} = \varnothing$ to emulate unconditional generation [138, 173, 237], thereby increasing the diversity of generated samples. However, in our context, a zero shape value corresponds to a plain body shape that lacks identity-specific signals.

We instead propose to perturb shape parameters slightly to augment the subject's identity, assuming that individuals with similar identities are likely to perform similar self-contact poses. As seen in Fig. 4.1, two persons performing the "rubbing belly" pose differently, particularly the right arm angles are non-identical. However, we observe that people with similar body shapes can come into contact with identical pose parameters. We therefore replace a normal shape conditioning $\boldsymbol{c} = \boldsymbol{I}$ with the perturbed shapes with a certain probability (*e.g.*, 30%) as

$$\boldsymbol{c} = \boldsymbol{I} + s_I \epsilon \tag{4.2}$$

where ϵ is standard Gaussian noise and s_I is a scaling factor to control the scale of the perturbation.

4.4.3 Training objectives

The training loss is computed by taking the difference between the original data $\boldsymbol{X}_0$ and the generated data $\hat{\boldsymbol{X}}_0$. We use the L1 loss between $\boldsymbol{X}_0$ and $\hat{\boldsymbol{X}}_0$ for the pose space, denoted as $\mathcal{L}_\theta$. We also compute losses on the mesh space after constructing meshes with the SMPL-X layer. We adopt the L1 loss between original and generated meshes for vertices $\mathcal{L}_v$, and an L1-based collision loss that penalizes vertices in collision on the generated mesh $\mathcal{L}_{col}$. The overall loss is formulated as

$$\mathcal{L}_D = \lambda_\theta L_\theta + \lambda_v L_v + \lambda_{col} L_{col}. \tag{4.3}$$

λ_θ, λ_v, and λ_{col} are the weights for each loss. We use 6D rotation representation [411] for pose parameters.

Collision detection: The collision loss $\mathcal{L}_{col}$ is designed to avoid heavy penetration on the mesh, which is essential to maintain plausible self-contact. For detection, [236] requires calculating pair-wise distances on vertices, but it is expensive to use

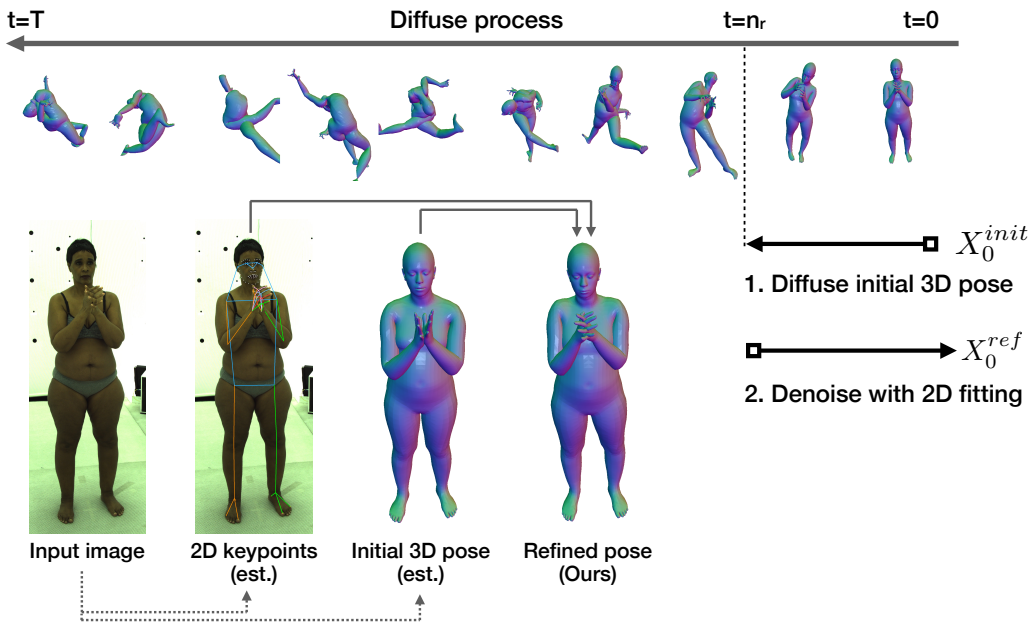


Figure 4.5: **Single-view refinement with diffusion.** Our refinement is based on the observations of 2D keypoints and initial 3D pose estimation. We diffuse the initial 3D pose $\boldsymbol{X}_0^{init}$ and then denoise it to obtain a refined pose $\boldsymbol{X}_0^{ref}$ while fitting to the 2D observation.

on the fly in training. This necessitates an efficient collision detector to consider collision in training.

We implement a fast ray-tracing-based collision detector following [311]. It casts rays in the normal direction from each vertex and computes the intersections with the mesh faces. Counting the number of intersections serves to find the vertices inside the mesh. Notably, vertices in the armpit region are often detected as collisions, leading to suboptimal solutions (*e.g.*, forcing the arms to move far from the torso). To address this, we restrict collision detection to areas relevant to hands, which are explicit targets in self-contact. Similarly to the data screening of Sec. 4.3, we apply the loss only to penetrating hand vertices and their corresponding vertices, focusing on hand-specific collisions such as hand-over-hand, hand-over-belly, and hand-over-face.

Algorithm 1 Single-view pose refinement: given initial 3D pose estimate $\boldsymbol{X}_0^{init}$, shape parameters $\boldsymbol{I}$, 2D keypoints to fit $\boldsymbol{P}_{2d}$, projection from pose to 2d keypoints M_{p2d}, a weight for 2d keypoint fitting λ_{2d}, start diffusion time n_r, mask for poses of interest $\boldsymbol{m}_p$.

▷ *Initialization: diffuse 3D pose at step* n_r

$\boldsymbol{X}_{n_r} \leftarrow \text{noise}\left(\boldsymbol{X}_0^{init}, n_r\right)$

for $n = n_r$ to 1 **do**

  $\hat{\boldsymbol{X}}_0 \leftarrow f\left(\boldsymbol{X}_n, n, \boldsymbol{I}\right)$

  ▷ *Optional: Blended pose denoising*

  $\hat{\boldsymbol{X}}_0 \leftarrow \hat{\boldsymbol{X}}_0 \odot \boldsymbol{m}_p + \boldsymbol{X}_0^{init} \odot (1 - \boldsymbol{m}_p)$

  $\epsilon_n \leftarrow \frac{1}{\sqrt{1-\bar{\alpha}_n}}\left(\boldsymbol{X}_n - \sqrt{\bar{\alpha}_n}\hat{\boldsymbol{X}}_0\right)$

  $\boldsymbol{X}'_{n-1} \leftarrow \sqrt{\bar{\alpha}_{n-1}}\hat{\boldsymbol{X}}_0 + \sqrt{1-\bar{\alpha}_{n-1}}\epsilon_n$

  ▷ *2D keypoint fitting*

  $\boldsymbol{X}_{n-1} \leftarrow \boldsymbol{X}'_{n-1} - \lambda_{2d}\nabla_{\boldsymbol{X}_n}\mathcal{L}_2\left(M_{p2d}(\hat{\boldsymbol{X}}_0), \boldsymbol{P}_{2d}\right)$

end for

return $\hat{\boldsymbol{X}}_0$

4.4.4 Inference

We describe data sampling in the following tasks. We use the DDIM sampling [314] for efficiency.

Random sampling: The trained diffusion model allows random data sampling via the reverse process from random noise (T: 1000 → 1). Inspired by [173], we reuse the collision loss in Sec. 4.4.3 as the additional guidance term, *anti-collision guidance*, which avoids the collision during the sampling phase as well. We set the sampling interval to 10. This is used to produce pose generation results of Sec. 4.5.2.

Single-view pose refinement: Observing self-contact poses inferred from single-view estimators [41, 180, 186, 227], the outputs often include incorrect contact states (*e.g.*, hands not in touch) due to the lack of contact prior, while detected 2D keypoints are aligned well with the given image. To address this, we develop single-view pose refinement, fitting the diffusion prior to the observed 2D keypoints with the estimated initial 3D poses. This does not require additional training compared to score distillation sampling [269] of [173], which is applicable to

any 2D/3D estimates.

Our refinement is efficient in sampling with fewer sampling steps (*e.g.*, only use the last 10% steps); see Algorithm 1. We assume given a single-view image, initial SMPL-X pose $\boldsymbol{X}_0^{init}$, shape $\boldsymbol{I}$, and 2D keypoints $\boldsymbol{P}_{2d}$ of the COCO-WholeBody format [154] can be estimated by off-the-shelf models, *e.g.*, recent vision foundation models like SMPLer-X [41] and Sapiens [158]. Then we sample data starting from the middle of the steps with diffused $\boldsymbol{X}_0^{init}$ at step n_r (*e.g.*, 100), reducing the number of sampling steps. In each iteration, we use the guidance of 2D keypoint fitting to $\boldsymbol{P}_{2d}$ by computing the gradient of the 2D keypoint error (L2 loss). We set the sampling interval to 1. This is used to produce the refinement results of Sec. 4.5.3.

We also provide an additional option of *blended pose denoising* in refinement. We find the 2D observation is likely to be partially unavailable or unreliable from in-the-wild videos, *e.g.*, upper-body videos in video conferences do not provide 2D keypoint cues for the lower body. Thus, inspired by image in-painting with diffusion models (*e.g.*, Blended Latent Diffusion [13]), we can only refine poses of interest (*e.g.*, upper body poses) during the reverse process, while the rest of the poses are unchanged. Specifically, given the mask for blending $\boldsymbol{m}_p$, we replace poses not to be refined with $\boldsymbol{X}_0^{init}$ in each iteration (Line 4 of Algorithm 1), ensuring the convergence to the initial poses. This simple trick helps control the inference process flexibly and enhances the applicability of the diffusion prior.

4.5 Experiments

We first present our dataset and implementation details in Sec. 4.5.1, and then provide results for pose generation with random sampling and single-view pose estimation with our refinement method in Secs. 4.5.2 and 4.5.3. We also show qualitative results in our proposed dataset.

4.5.1 Experiment setup

Datasets: We create *train*/*eval* sets in the Goliath-SC dataset. The *train* set is constructed with the captures with action instructions (*e.g.*, rubbing neck), which comprises 313K poses. Additionally, the *eval* set is designed for single-view pose

Method	FID↓	KID↓ ($\times 10^{-3}$)	Div.↑	Prec.↑	Recall↑	Col. ratio↓
			Unconditional generation			
VPoser* [266]	9.43	0.930	3.34	1.0	0.006	1.72
BUDDI* [237]	3.19	1.16	6.36	0.957	0.528	1.32
			Shape-conditional generation			
VPoser* [266]	9.16	0.882	3.20	**1.0**	0.005	**1.37**
BUDDI* [237]	2.66	1.12	5.59	0.995	0.488	1.47
Ours	**1.25**	**0.430**	**5.98**	0.985	**0.708**	1.52

Table 4.2: **Results of self-contact pose generation.** We study sample quality and diversity in generation without (unconditional) or with shape conditioning, evaluated on the *train* split. The notation * indicates the methods adapted to our task.

estimation, featuring *unseen subjects*, which contains 9.7K samples. This is used to test generalizability to unseen subjects where the same action instructions are given as the *train* set.

Implementation details: For generation, we use two self-attention layers with latent_size=256, depth=4, num_heads=4, and set λ_θ=1, λ_v=1e-3, λ_{col}=1e-4. We set N_s to the full size of 300 to incorporate as much detailed shape information as possible. We set the shape perturbation probability and s_I to 0.3 and 1e-4. Following [173, 278, 331], we report Fréchet Inception Distance (FID) [134], Kernel Inception Distance (KID) [28], diversity, and precision-recall [297] for the evaluation. We also show the collision ratio of collided vertices over all SMPL-X vertices, using the detector of $\mathcal{L}_{col}$ in Sec. 4.4.3.

For single-view pose estimation, we prepare different SMPL-X regressors, namely HybrIK-X [180, 186], Hand4Whole [227], and SMPLer-X [41], and use Sapiens [158] for 2D keypoint detection. We report the MPJPE for 3D keypoints of the COCO-WholeBody format [154] in the body-root aligned coordinates (disregarding global rotation and translation). We set λ_{2d} and n_r to 0.01 and 100.

Method	FID↓	Div.↑	Col. ratio↓
w/o Shape cond.	2.18	5.52	1.92
w/o PASA	1.42	5.74	1.62
w/o Shape rand.	1.27	5.89	**1.41**
w/o Anti-col.	1.28	**6.01**	1.85
Ours	**1.25**	5.98	1.52

Table 4.3: **Ablation study in our generation.** We compare methods without shape conditioning (Shape cond.), part-aware self-attention (PASA), shape perturbation (Shape rand.), and anti-collision guidance (Anti-col.).

Figure 4.6: **Qualitative results of our generation with shape interpolation**. We interpolate between two shape parameters with the fixed latent code (*i.e.*, starting with the same noise at $t = T$). Our model generates plausible self-contact poses under varying shapes.

4.5.2 Pose generation

Tab. 4.2 shows pose generation results with or without shape conditioning in Goliath-SC. We compare our diffusion method with VAE-based VPoser [266] and diffusion-based BUDDI [237]. We modify these baselines to our task to take the whole body pose parameters as input (aligned to $\boldsymbol{X}$ of Sec. 4.4.2), denoted as VPoser* and BUDDI*.

The results show the superiority of our shape-conditional method, remarked by the improvement over the VAE and diffusion baselines. While VPoser* easily overfits to higher precision, our diffusion method (Ours) has significantly improved recall with a smaller FID score. Our method further surpasses BUDDI*, a

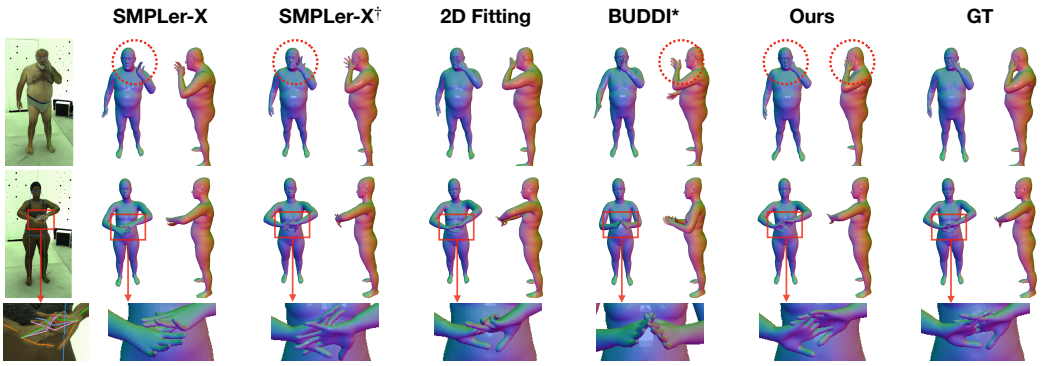


Figure 4.7: **Qualitative results of our single-view refinement on Goliath-SC.** Our method successfully refines the initial poses to be valid self-contact for fine-grained poses, such as face touching and two-hand overlap.

method without latent diffusion and part-wise attention, achieving a 53% reduction in FID. Since random outputs exhibit higher diversity and samples lacking contact yield lower col. ratio, maintaining higher diversity and lower col. ratio scores with improved FID and KID (smaller distribution distance) is vital; our method shows a better balance between quality and variety.

Ablation study: Tab. 4.3 shows the ablation study of our proposed method. Shape-conditional generation surpasses unconditional baselines (w/o Shape cond.), which is also observed in the VPoser* and BUDDI* of Tab. 4.2. This indicates that the body shapes help to learn self-contact pose distribution effectively. The ablation study shows that part-aware self-attention (PASA), shape perturbation (Shape rand.), and anti-collision guidance (Anti-col.) reduce FID scores consistently over those without each module. Specifically, anti-collision guidance helps reduce col. ratio in test time with improved FID.

Qualitative results: Fig. 4.6 shows our qualitative results with shape interpolation. When changing the body shapes (from a large to a slim body), the generated poses continuously move on the hand surface while preserving plausible self-contact poses. This indicates that our diffusion model can learn a smooth manifold of self-contact poses with respect to body shape changes.

Method	Avg.	Hands	Body	Face
Hand4Whole [227]	126.3	225.8	89.6	78.0
+ 2D fitting	89.5	179.9	64.8	47.1
+ BUDDI* [237]	74.5	109.2	37.8	65.9
+ Ours (w/o Shape cond.)	37.9	74.6	29.4	**18.2**
+ Ours	**35.3**	**66.6**	**26.5**	18.3
HybrIK-X [180]	82.3	99.2	62.8	76.2
+ 2D fitting	51.8	63.4	38.4	50.5
+ BUDDI* [237]	65.0	90.5	36.0	58.8
+ Ours (w/o Shape cond.)	45.9	85.5	32.5	26.3
+ Ours	**32.4**	**58.7**	**26.1**	**17.5**
SMPLer-X [41]	58.0	98.7	41.6	38.9
SMPLer-X†	42.0	56.7	31.9	34.1
+ 2D fitting	41.7	65.7	30.6	31.6
+ BUDDI* [237]	71.7	99.9	36.3	66.4
+ Ours (w/o Shape cond.)	33.7	63.6	26.1	**17.4**
+ Ours	**31.8**	**54.6**	**24.7**	19.2

Table 4.4: **Results of single-view pose regression in Goliath-SC.** We evaluate our diffusion-based pose refinement in the *eval* set given initial pose estimation from SMPL-X regressors. We report MPJPE in millimeter on the body-root aligned coordinates. The notation † shows fine-tuned results for the dataset.

4.5.3 Single-view pose estimation and refinement

Analysis on SMPL-X regressors: Tab. 4.4 shows single-view pose estimation results in Goliath-SC. We first evaluate existing SMPL-X regressors [41, 180, 227] in self-contact scenarios. Hand4Whole and HybriIK-X adopt CNN-based backbones (*i.e.*, ResNet [131] and HRNet [322]). SMPLer-X [41] is a foundation model trained on self-contact datasets (*i.e.*, MTP [236] and HumanSC3D [94]). We further fine-tune the model on Goliath-SC, denoted as SMPLer-X†.

In the predictions from Hand4Whole and HybriIK-X, we find frequent failures in handling to place hands in contact, *i.e.*, 2D pose is aligned in the image view but higher depth errors are present for hands. Owing to the higher diversity in the training data, SMPLer-X facilitates tracking better poses in our Goliath-SC data, with 58.0 mm error (see Fig. 4.7). We confirm the state-of-the-art performance in image-based regression with the fine-tuned SMPLer-X†, exhibiting an overall error reduction to 42.0 mm.

Analysis on refinement: We compare our proposed refinement (Algorithm 1) with conventional refinement. This setup assumes 2D/3D observations are given, namely initial SMPL-X estimates and 2D keypoints. A naive baseline is a simple 2D keypoint fitting with optimization, which is widely used for pseudo-mesh registration of SMPL-X on in-the-wild videos [140, 192, 228]. Since the body and face can be well-constrained with 2D keypoints, their gains are better than the fine-tuned results in the SMPLer-X setting. In contrast, fitting to hands causes implausible 3D hand poses though the 2D projection error is minimized. This underscores the need for the model prior in self-contact, particularly to correct hand placement and its local pose.

Next, we evaluate our refinement with diffusion-based priors, namely, our PA-PoseDiff, and BUDDI* trained in Tab. 4.2. BUDDI*, a method without latent diffusion and part-wise self-attention, shows effectiveness when the initial estimation is noisy (Hand4Whole and HybrIK-X), while it has limited refinement capacity when the initialization is reasonable (SMPLer-X). This suggests that additional refinement of well-estimated poses requires more precise modeling of self-contact poses, which is essential to achieve improvements beyond the initial quality.

Our final diffusion prior achieves significantly improved results by reducing

overall and part-wise errors across the three settings with different regressors. Our refinement demonstrates less dependency on the initialization and stable performance, as post-refinement converges to lower errors with the varied initializations. Compared to the unconditional prior of our method (w/o Shape cond.), the proposed shape-conditional prior achieves better results, particularly for hands and body. This indicates that the shape-dependent constraint is effective in capturing the relationship between the body and hands, as they have a higher correlation with body shapes than the face.

Qualitative results: Fig. 4.7 shows qualitative results of our refinement. We find that the initial predictions include ambiguities in contact and depth estimation, *e.g.*, interacting parts are not in contact, especially for fine details, and high-depth errors remain for hands. Our method can correct such failures with the generative prior derived from the contact data only, without requiring knowledge of where to contact.

Discussion: regression *vs*. generative prior: While recent 3D pose estimators are trained on extensive human data, the state-of-the-art baseline with regression still struggles to estimate self-contact poses of unseen subjects. In contrast, our approach introduces a novel *generative prior modeling* of self-contact pose distribution, with the body-shape dependent assumption. Our model not only generalizes better to new subjects but also enhances robustness in handling fine-grained self-contact poses. This suggests that our generative prior offers a flexible and scalable solution for self-contact modeling over the regression approach.

4.6 Additional Results and Details

4.6.1 Dataset details

Our dataset is constructed on the minimally-clothed body setup of [215], which aligns with the previous work on body shape prior captured with subjects in tight-fit clothing [206, 266]. We obtain user consent for data captures and the release of the registered SMPL-X parameters.

Due to the fixed capture space, most samples do not have high variations for lower-body poses. Nevertheless, this setup allows for capturing *intricate* upper-body self-contact details (*e.g.*, "rubbing eyes") with unprecedented fidelity unavailable in existing studies [93, 94, 236, 385]. While modeling large variations in lower-body pose (*e.g.*, tying shoelaces) is not prioritized in this work, we will consider an expanded capture setup as future work.

SMPL-X registration: We initiate with a human mesh model used in [215] that has a uniform topology across subjects, and we pre-compute its vertex-face correspondence to SMPL-X using barycentric coordinates. We register the human model across frames while tracking pose and surface precisely without relying on mocap markers. Given multi-view dome captures, we first fit the human model to the rest pose (A-Pose). Then we run 3D pose tracking based on multi-view images over the frames and use Linear Blend Skinning (LBS) that transforms a mesh in the rest pose to the desired pose of each frame. The subject' s poses are continuously captured at 30 Hz with scripted action instructions to let participants express the corresponding gestures. Given the registered mesh, the SMPL-X registration is obtained through vertex-to-vertex alignment between two meshes[1], as shown in Fig. 4.8. The continuous poses in our capture allow stable mesh alignment by using the previous frame's registration as initialization for the current frame, preventing significant fitting failures.

Data statistics: Fig. 4.9 details action scripts used in the capture and the number of self-contact poses per action. The nouns of the actions suggest interacting body parts as following groups.

- Head-related: Face, forehead, temples, eyes, nose, hair, facial hair, and neck

[1] https://github.com/vchoutas/smplx/blob/main/transfer_model/docs/transfer.md

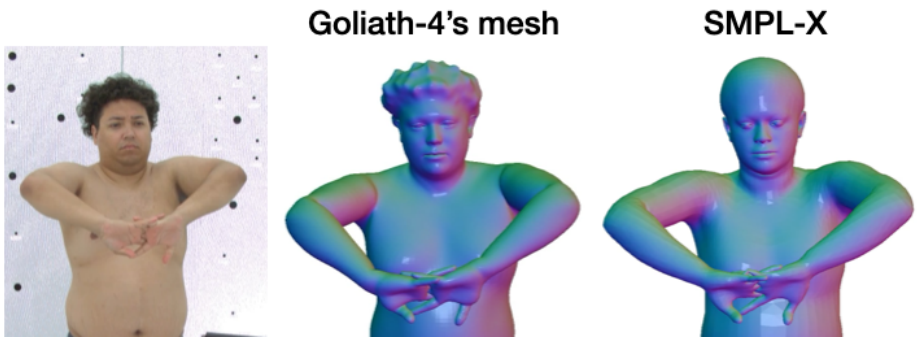


Figure 4.8: **Conversion from Goliath-4 [215]'s mesh to SMPL-X**.

- Upper body-related: Arm, hand, wrist, fingers, thumb, and palm
- Torso-related: Belly, back, lumbar, and thighs

Instead, the verbs indicate how to interact with the body part; general movements are represented, such as hitting, grabbing, holding, clapping, rubbing, massaging, scratching, punching, wrapping, and itching. In addition, hand-specific movements include extension, flex, rotation, press, snap, interlock, touch, and squeeze. These hand interactions tend to be in close contact mostly in the captured sequence, resulting in a large number of poses in self-contact, such as "hands massaging hands".

Our dataset is constructed by capturing 130 subjects where the gender distribution is detailed in Tab. 4.1. To confirm the variety of captured shape information, we provide comprehensive analysis on shape statistics in Fig. 4.10, *i.e.*, standard deviation and range of 10 shape components of SMPL-X compared to the existing self-contact datasets, such as HumanSC3D [94] and MTP [236]. Our dataset (red) has the largest variety in most components except for 6th range, 8th std and range, indicating higher subject diversity and variability of our Goliath-SC.

4.6.2 Additional implementation details

Baselines: We detail the implementation of baselines used in our experiments. For fair comparison, we retrain the comparison models from scratch with the same input representation of the whole-body pose parameters (aligned to $\boldsymbol{X}$ of Sec. 4.4.2 , including hands, body, and face.

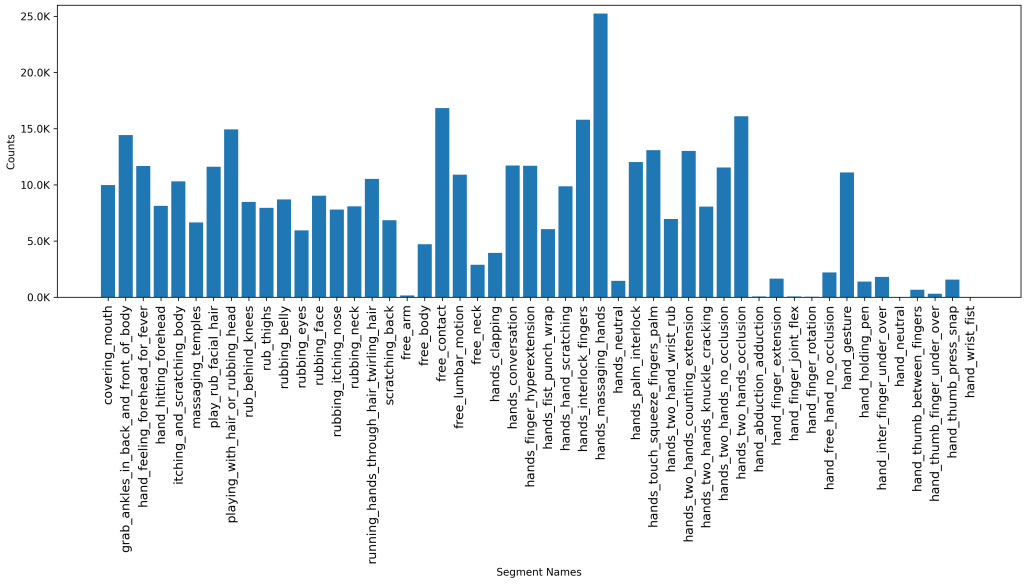


Figure 4.9: **Statistics of scripted actions and the number of self-contact poses in Goliath-SC.**

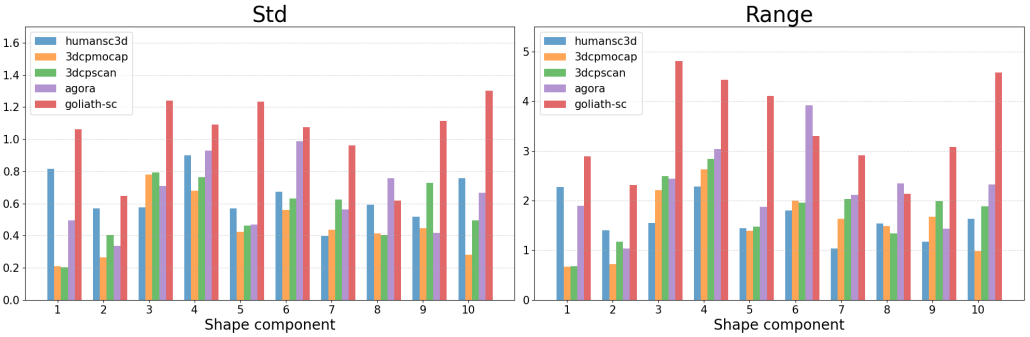


Figure 4.10: **Variability of subject shapes.** Standard deviation and range (max − min) of the first 10 shape components in self-contact datasets, namely HumanSC3D [94], (3DCPMocap, 3DCPScan, Agora) from MTP [236], and our Golaith-SC.

BUDDI [237] is originally proposed for two-body interactions, modeling the joint distribution of the body pose parameters of SMPL-X (compatible to SMPL) and its shape parameters without latent diffusion modeling. A two-hand interaction generation model, InterHandGen [173], shares a similar architecture. Our implemented BUDDI* modifies the original BUDDI to take the whole-body pose parameters of a single person. Following the original implementation, the transformer layers are used and all pose parameters are concatenated to a single vector, which indicates the absence of part-wise attention compared to our PAPoseDiff. To produce the results of Tab. 4.2 , we construct baselines of the joint distribution modeling between pose and shape, *i.e.*, unconditional model, and shape-conditional pose generation with the input embedding of the shape parameters. In addition, when adapting to the task of single-view refinement, we use our fitting algorithm (Algorithm 1) with the unconditional BUDDI* prior. The hyper-parameters in the fitting (*e.g.*, weight for 2D keypoint fitting and start diffusion time) follow those of our final method.

VPoser [266] is a VAE-based pose prior that learns pose distribution on the body pose parameters of SMPL-X. Similarly to BUDDI*, we adapt this architecture for our task, by taking the whole-body pose parameters and adding shape parameters as conditional input, denoted as VPoser*.

Training details: We train generative models for 150,000 iterations with a batch size of 32, using an Adam optimizer [163] with a learning rate of 1e-4. Our diffusion process is based on cosine noise scheduling with T=1000. Unlike the conventional choice of using 10 shape components of SMPL-X, we input the full 300-dimensional vector of the shape parameters to let the generative prior access as much fine details as possible (*e.g.*, hands and face shapes).

4.6.3 Additional results

Qualitative results of single-view pose estimation: Additional qualitative results of single-view pose estimation are found in Fig. 4.11, including SMPLer-X, fine-tuned SMPLer-X†, 2D fitting, BUDDI*, our final refinement (Ours), and GTs. We observe that hands are not often in contact with SMPLer-X (*e.g.*, Rows 1,2,4,6), while the fine-tuned baseline struggles with highly bent hand fingers (*e.g.*, Rows 3,6,7) and incorrect contact states, *e.g.*, for the hidden left hand behind the neck of

Row 5. The 2D keypoint fitting baseline tends to exhibit unsolved depth ambiguity (Rows 1,6) and implausible hand poses (Row 1) due to overfitting to the 2D observation. The BUDDI* method often relies heavily on the model prior with a large 2D error to the observation. This indicates that the method generates plausible poses, yet not aligned to the given 2D keypoints, such as Rows 1,2,3,5. It also comprises higher hand depth errors (to those to be in contact) like Rows 1,2,6. Notably, our method can resolve these failures presented by the comparison models and shows significantly reduced errors in 3D compared to the GT.

The last row shows a remaining failure when fingers are in complex interaction, *i.e.*, the fingers of both hands, except for the ring fingers, are overlapped while only the ring fingers are bent. Neither method handles this pose well because of the inaccuracy of 2D keypoint detection. Improving detection and model-based refinement to such fine interactions are future challenges.

Figure 4.11: **Qualitative results of single-view pose estimation.** The four subjects of Goliath-4 [215] are illustrated.

4.7 Conclusion

To highlight challenging self-contact scenarios, we offer a comprehensive self-contact analysis, along with the newly captured **Goliath-SC** dataset with 383K poses and precise body shape registration. We then model the self-contact pose manifold depending on body shapes with the generative diffusion model. Specifically, the latent diffusion with part-aware self-attention helps to learn pose distribution effectively and achieves the best in pose generation. We further propose single-view pose refinement using the diffusion prior, while fitting to the observed 2D keypoints. Our experiments confirm the successful refinement of self-contact poses and show our superiority over the state-of-the-art diffusion method and the regressive foundation model.

Limitation and future work: We observe that hand-hand interaction is still a challenging scenario in the generation, in which minor interpenetration persists due to higher articulation, as studied in [173]. In addition, not only addressing in-contact scenarios only, but also generalizable modeling to non-contact cases like [236] is an interesting extension. Our success in self-contact modeling opens new avenues to include additional self-contact scenarios, *e.g.*, without scripted action instructions, with various global body poses (*e.g.*, sitting), or in multi-person conversation. Extending the diffusion prior into the temporal dimension or with linguistic contents is promising future work.

Chapter 5

3D Hand Pose Pre-training from In-the-wild Images

In this chapter, we present a framework for pre-training of 3D hand pose estimation from in-the-wild hand images sharing with similar hand characteristics, dubbed **SiMHand**. Pre-training with large-scale images achieves promising results in various tasks, but prior methods for 3D hand pose pre-training have not fully utilized the potential of diverse hand images accessible from in-the-wild videos. To facilitate scalable pre-training, we first prepare an extensive pool of hand images from in-the-wild videos and design our pre-training method with contrastive learning. Specifically, we collect over 2.0M hand images from recent human-centric videos, such as *100DOH* and *Ego4D*. To extract discriminative information from these images, we focus on the *similarity* of hands: pairs of non-identical samples with similar hand poses. We then propose a novel contrastive learning method that embeds similar hand pairs closer in the feature space. Our method not only learns from similar samples but also adaptively weights the contrastive learning loss based on inter-sample distance, leading to additional performance gains. Our experiments demonstrate that our method outperforms conventional contrastive learning approaches that produce positive pairs solely from a single image with data augmentation. We achieve significant improvements over the state-of-the-art method (PeCLR) in various datasets, with gains of 15% on FreiHand, 10% on DexYCB, and 4% on AssemblyHands.

5.1 Introduction

Hands serve as a trigger for us to interact with the world, as seen in various human-centric videos. The precise tracking of hand states, such as 3D keypoints, is crucial for video understanding [303, 359], AR/VR interfaces [119, 362], and robot learning [49, 276]. To this end, 3D hand pose estimation has been studied through constructing labeled datasets [49, 251, 252, 421] and advancing supervised pose estimators [40, 89, 104, 198, 265]. However, utilizing large-scale, unannotated hand videos for pre-training remains underexplored, while collections of human-centric videos, like 3,670 hours of videos from Ego4D [111] and 131-day videos from 100DOH [305], are readily available.

In pre-training, contrastive learning has been utilized to learn from unlabeled images like SimCLR [59], which maximizes agreement between positive pairs while repelling negatives. Spurr *et al*. [316] introduce pose equivariant contrastive learning (PeCLR) for 3D hand pose estimation, which aligns the geometry of features encoded from augmented images with affine transformations. However, both SimCLR and PeCLR create positive pairs from a single sample by applying data augmentation, limiting the gains from positive pairs as their hand appearance and backgrounds are identical. Ziani *et al*. [418] extend the contrastive learning framework to video sequences by treating temporally adjacent hand crops as positive pairs. However, in-the-wild videos can challenge tracking hands across frames, especially in egocentric views where hands are often unobservable due to camera motion. Meanwhile, this temporal positive sample mining remains the limited appearance variation of hands and backgrounds.

In this work, we introduce SiMHand, a novel contrastive learning framework for 3D hand pose pre-training, which leverages diverse hand images in the wild, with the largest 3D hand pose pre-training set to date. We specifically collect 2.0M hand images from human-centric videos, from Ego4D [111] and 100DOH [305], using an off-the-shelf hand detector [305]. Our pre-training set significantly exceeds the scale of prior works by two orders of magnitude, such as over 32-47K images in [316] and 86K images from 100DOH in [418].

Our method focuses on learning discriminative information by mining hands with similar characteristics from various video domains. Based on our observations, contrastive learning can further benefit from discriminating the foreground

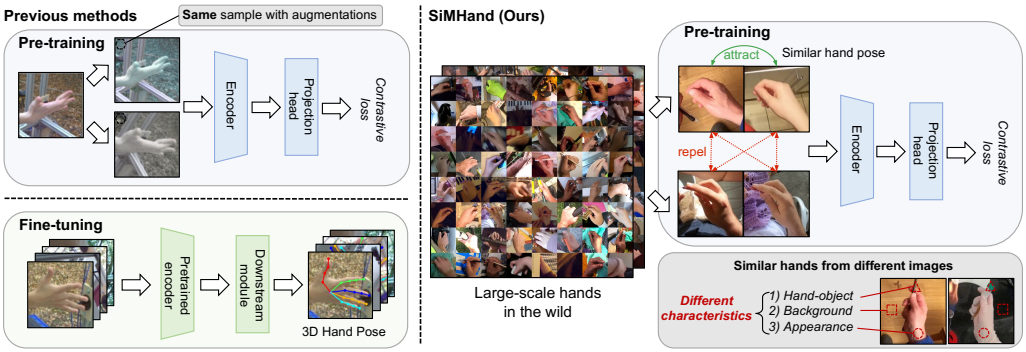


Figure 5.1: **The pipeline of pre-training and fine-tuning. (Left)** Previous pre-training methods (*e.g.*, PeCLR [316]) learn from positive pairs originating from the different augmentations and fine-tune the network on a dataset. **(Right)** Our method is designed to learn from positive pairs with similar foreground hands, sampled from a pool of hand images in the wild.

of hands in varying backgrounds. As shown in Fig. 5.1, our positive pairs are sourced from different images, offering additional information gains from different types of object interactions, backgrounds, and hand appearances. Specifically, we use an off-the-shelf 2D hand pose estimator [209] to identify similar hands from the pre-training set.

Using the identified similar hands as positive pairs, we further propose adaptive weighting, to dynamically find informative pairs during training. A naive adaptation of the similar hands is to replace the original positive pairs in contrastive learning, but this scheme struggles to exploit *how similar the paired hands are*. To tackle this, we assign weights based on the similarity scores within the mini-batch in the contrastive learning loss. The weights are designed to have higher values as the similarity of the pairs increases. This allows the optimization of contrastive learning to explicitly consider the proximity of samples, beyond binary discrimination between positives and negatives.

We validate the effectiveness of the pre-trained networks by fine-tuning on several datasets for 3D hand pose estimation, namely FreiHand [421], DexYCB [49], and AssemblyHands [252]. Our proposed method consistently outperforms conventional contrastive learning methods, SimCLR and PeCLR. Additionally, we conduct extensive ablation experiments to analyze: 1) performance with varying

pre-training and fine-tuning data sizes, 2) the effect of adaptive weighting, and 3) the improvement with different levels of similarity.

In summary, the main contribution of this paper is threefold:

- We propose SiMHand, a contrastive learning method for 3D hand pose pre-training, leveraging positive samples with similar hands mined from 2.0M in-the-wild hand images.
- We introduce a parameter-free adaptive weighting mechanism in the contrastive learning loss, enabling optimization guidance according to the calculated similarity.
- Our experiments demonstrate that our approach surpasses prior pre-training methods and achieves robust performances across different hand pose datasets.

5.2 Related Work

3D hand pose estimation: The task of 3D hand pose estimation aims to regress 3D hand joints. Since annotating 3D hand poses is challenging, only limited labeled datasets are available [251], and most of which are constructed in controlled laboratory settings [49, 230, 252, 421]. Given this challenge, two approaches have been proposed to facilitate learning from limited annotations: pseudo-labeling and self-supervised pre-training. Pseudo-labeling methods learn from pseudo-ground-truth assigned on unlabeled images [65, 198, 200, 255, 377, 405]. For example, S2Hand [65] attempts to learn 3D pose only from noisy 2D keypoints on a single-view image, while HaMuCo [405] extends such self-supervised learning to multi-view setups. Alternatively, pre-training methods aim to find well-initialized models with unlabeled data for downstream tasks. Prior works propose contrastive learning approaches but rely on relatively small pre-training sets (*e.g.*, 32-47K images in [316] and 86K images in [418]). We collect hand images from large human-centric datasets such as Ego4D [111] and 100DOH [305], expanding our pre-training set to 2.0M images.

Contrastive learning: Contrastive learning has emerged as a powerful technique in self-supervised learning, bringing positive samples closer while pushing negative samples apart [71, 130, 146, 302, 312, 313]. Standard methods

generate positive samples from an identical image with data augmentation (*i.e.*, self-positives) [46, 47, 60, 113, 280], thus the positive supervision doesn't explicitly model inter-sample relationships. To address this, Zhang *et al*. propose a relaxed extension of self-positives, *non-self-positives* [400], which share similar characteristics but originate different images, such as images capturing the same scene [9, 25, 106, 128], the same person ID [53, 68], and multi-view images [153]. The positive supervision from non-self-positives enables considering diverse inter-sample alignment and facilitates the learning of semantics more easily. Zhang *et al*. identify non-self-positives by searching similar human skeletons from single-view images and adapt in action recognition [400]. Jie *et al*. rely on multi-view (*i.e*. paired) images to define non-self-positives and propose pair-wise weights to adaptively leverage useful multi-view pairs [153]. Our work proposes the mining of non-self-positives from 2D keypoint cues with additional pair-wise weighting to account for similarity from *unpaired* data in pre-training.

5.3 Method

Our approach SiMHand aims to pre-train an encoder for 3D hand pose estimation with large-scale human-centric videos in the wild. We first construct a pre-training set from egocentric and exocentric hand videos (Sec. 5.3.1). Then, we find similar hand images to define positive pairs across videos (Sec. 5.3.2). Finally, we incorporate these positive pairs into a contrastive learning framework and employ adaptive weights to improve the effectiveness in pre-training (Sec. 5.3.3).

5.3.1 Data preprocessing

Our preprocessing involves creating a set of valid hand images for pre-training, which is sampled from a set of N videos: $\{v_1, v_2, \ldots, v_N\}$. We use an off-the-shelf hand detector [305] to select valid frames with visible hands. Given a video frame $I_{\text{full}} \in v_i$, the model detects the existence of the hand and its bounding box, creating hand crops enclosing either hand identity (right/left) from I_{full}. To avoid bias related to hand identity, we balance the number of right and left hand crops equally and then convert all crops to right-hand images. Then, we create a set of frames for each video v_i as $\mathcal{F}_i = \{I_{i,1}, I_{i,2}, \ldots, I_{i,T_i}\}$, where $I_{i,j} \in \mathbb{R}^{H \times W \times 3}$ rep-

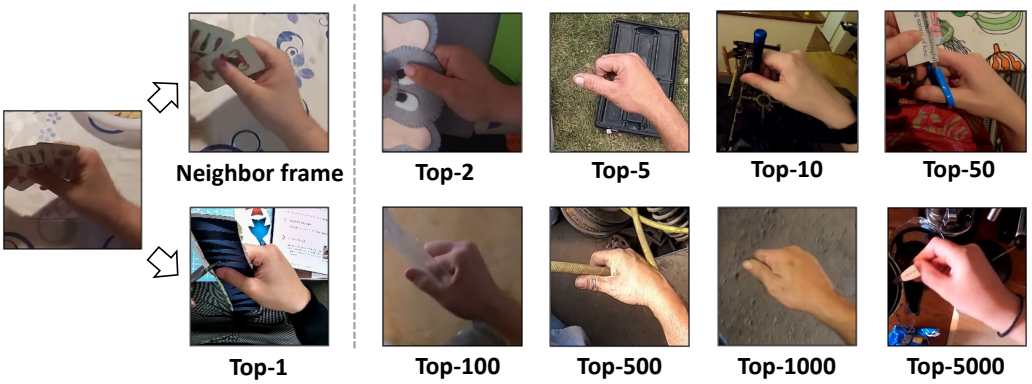


Figure 5.2: **Visualization of similar hand samples in Top-K.** Given the query image (I), the mined similar samples are shown ("Top-1" corresponds to I^+ in Sec. 5.3.2).

resents the processed crop with height H and width W, and T_i is the total number of crops in v_i. The height H and width W are defined post-resize to give the uniform image size. Using this frame set $\mathcal{F}_i$, the video dataset can be re-represented as $\mathcal{V} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N\}$. Specifically, we processed two datasets, Ego4D [111] and 100DOH [305], to collect 1.0M images from 8K and 21K videos, respectively.

5.3.2 Mining similar hands

To incorporate diverse samples in contrastive learning, we design positive pairs from non-identical images with similar foreground hands. Here we construct a mining algorithm to find similar hands from $\mathcal{V}$ by focusing on pose similarity between hand images. We first extract 2D keypoints from I, embed in the feature space, and search a positive sample.

Pose embedding: We adopt estimated 2D keypoints (for 21 joints) to find similar hands. We use an off-the-shelf 2D hand pose estimator ϕ [209], but the outputs are prone to be noisy in testing in the wild. To make it more robust, we obtain a D-dimensional embedding of 2D hand keypoints, $\mathbf{p} \in \mathbb{R}^D$, for each image I. This serves to reduce the noise effect while preserving the semantics of hands. We use a concatenated 42-dimensional vector as the output of ϕ for later use. Particularly, we apply PCA-based dimension reduction, which projects the keypoints vector

into a lower-dimensional space of size D. Given the PCA projection matrix $M \in \mathbb{R}^{42\times D}$, the pose embedding $\mathbf{p}$ is calculated as $\mathbf{p} = M^T\phi(I)$.

Mining: This step is designed to identify a positive sample $I^+ \in \mathbb{R}^{H\times W\times 3}$ paired with a query image I. We denote the similarity mining logic as $I^+ = \mathrm{SiM}(I)$. As shown in Fig. 5.2, using the closest (neighbor) sample in the PCA space encounters a trivial solution $I, I^+ \in v_i$, where both images originate from the same video v_i. Similarly to [418], the supervision by neighbor samples of the same video has less diversity in backgrounds, hand appearances, and object interactions. Thus we are motivated to find similar hands derived from different videos. Specifically, we search the minimum distance within the set of all frames except for v_i, written as $\mathcal{F}_i^c = \bigcup_{k\neq i}\mathcal{F}_k$. Given an query $I_{i,j}$, which represents the j-th image of the i-th video, the function $\mathrm{SiM}(\cdot)$ is formulated as

$$\mathrm{SiM}(I_{i,j}) = \arg\min_{x\in\mathcal{F}_i^c} D(M^T\phi(x), M^T\phi(I_{i,j})), \tag{5.1}$$

where $D(\cdot,\cdot)$ is the Euclidean distance metric.

As a proof of concept, we illustrate examples after our mining $\mathrm{SiM}(\cdot)$ in Fig. 5.2. We denote "Top-1" (most similar) as our assigned positive sample I^+ to the query image I. As references, the rest of the figures ("Top-K") represent the K-th similar samples. Our sampling highlights the diversity in captured environments and interactions, while it also suggests that as the rank (distance) increases, the sampled images become dissimilar.

5.3.3 Contrastive learning from similar hands with adaptive weighting

We detail our contrastive learning approach (see Fig. 5.3), learning from mined similar hands with adaptive weighting.

Overview: The contrastive learning is designed to align positive samples (I, I^+) in the feature space, constructed in Sec. 5.3.2, and the rest of negative samples are pushed apart. Following [59, 316], we treat all mini-batch samples other than the corresponding positive samples as negative samples I^-. Feature extraction is performed by two learnable components: an encoder $E(\cdot)$ and a projection head $g(\cdot)$, which indicates the entire model as $f = g \circ E$. The extraction is combined with image augmentation $\mathbf{T}$, which formulated as $\mathbf{z} = f(\mathbf{T}(I))$ and $\mathbf{z}^+ = f(\mathbf{T}(I^+))$.

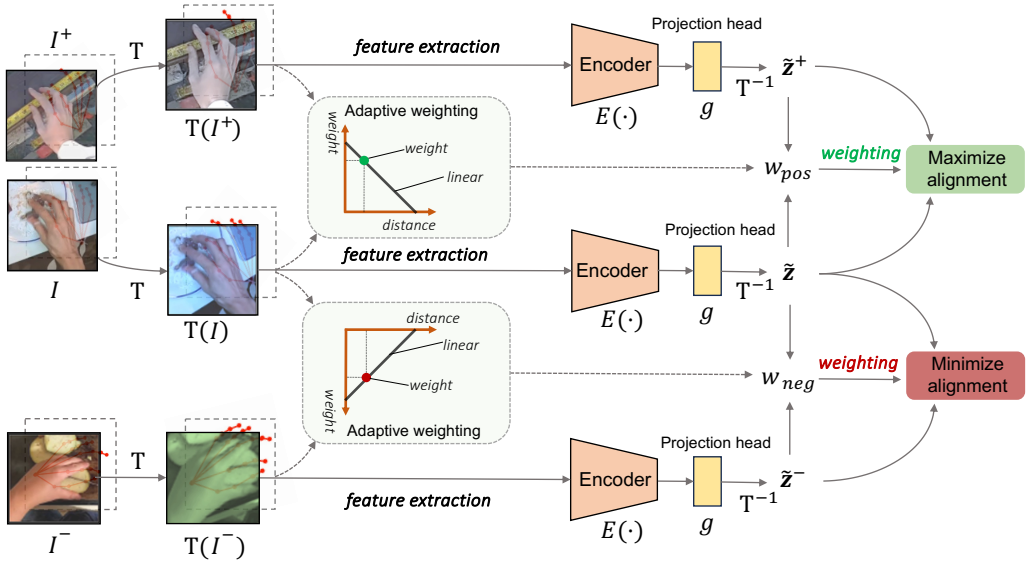


Figure 5.3: **Overview of our SiMHand.** Starting from the left, hand images (I, I^+, I^-) and their corresponding 2D keypoints are input to the model. After applying random augmentations through transformation $\mathbf{T}$, both the images and 2D keypoints are spatially transformed. The altered 2D keypoints are then used to compute adaptive weights w_{pos} and w_{neg}, which guide contrastive learning by strengthening or weakening the alignment between positive and negative samples.

Applying geometric transformations (*e.g.*, rotation) in $\mathbf{T}$ can cause misalignment between the image and feature spaces; we correct such an error with the inverse transformation $\mathbf{T}^{-1}$ as [316]. After applying the inverse transformation to the feature $\mathbf{z}$, we obtain a feature $\tilde{\mathbf{z}} = \mathbf{T}^{-1}(\mathbf{z})$, where geometry is aligned to the original images. As such, all anchor, positive, and negative samples are encoded as $\tilde{\mathbf{z}}$, $\tilde{\mathbf{z}}^+$, and $\tilde{\mathbf{z}}^-$, respectively.

Adaptive weighting: During learning from our similar hands, we propose an adaptive weighting per pair to focus more on informative samples that provide greater discriminative information. The assigned weights are computed by the predefined similarity metric in Sec. 5.3.2. Given pre-processed keypoints for two samples within the mini-batch, $\mathbf{k}_1$, $\mathbf{k}_2$, the weight w is computed by linear scaling with the Euclidean metric $D(\cdot, \cdot)$ as

$$w = \frac{d_{\text{max}} - D(\mathbf{k}_1, \mathbf{k}_2)}{d_{\text{max}} - d_{\text{min}}}, \tag{5.2}$$

where d_{min}, d_{max} are the minimum and maximum distances within the mini-batch. This assigned weight w dynamically changes according to the sample statistics in the mini-batch, enabling adaptive attention per iteration.

To address the distinction between positive and negative sample weighting, we introduce separate weighting terms for positive and negative pairs. Specifically, w_{pos} corresponds to the weight assigned to positive pairs, while w_{neg} is used for positive-negative pairs.

Contrastive loss with weighting: We finally formulate contrastive learning with the proposed weighting scheme. We assume that a mini-batch contains 2N samples in total, including N query samples and their corresponding N positive samples. We introduce separate weighting terms for positives (I, I^+) and negatives (I, I^-) as w_{pos} and w_{neg}, respectively. With these weights, our constrastive learning loss based on the NT-Xent loss [59] is formulated as:

$$\mathcal{L}_i = -\log \frac{\exp\left(w_{pos} \cdot \text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_i^+)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(w_{neg} \cdot \text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_k^-)/\tau\right)} \tag{5.3}$$

Here τ is a temperature parameter, $\text{sim}(\mathbf{z}, \bar{\mathbf{z}}) = \frac{\mathbf{z}^T \bar{\mathbf{z}}}{\|\mathbf{z}\| \|\bar{\mathbf{z}}\|}$ is the cosine similarity function. Overall, our adaptive weighting enables considering the importance separately for positive and negative samples, while closer samples are assigned with higher weights and more distant ones receive lower weights.

5.4 Experiments

In this section, we compare our method with existing baselines for pre-training of the 3D hand pose estimation and conduct ablation experiments to support the validity of our approach. We begin by providing a detailed explanation of the dataset and experimental setup (Sec. 5.4.1). Next, we demonstrate that our model achieves competitive performance compared with existing methods (Sec. 5.4.2). Following this, we present the results of ablation studies on weighting design in the pre-training phase (Sec. 5.4.3). Finally, visualizations are used to illustrate the superiority and efficiency of our approach (Sec. 5.4.4).

5.4.1 Experimental setup

Pre-training datasets: We curate a large collection of hand images from two major video datasets, Ego4D [111] and 100DOH [305], featuring egocentric and exocentric views respectively. From Ego4D, a vast egocentric video dataset with 3,670 hours of footage, we extracted 1.0M hand images from 8K videos. Similarly, from the exocentric dataset 100DOH, which includes 131 days of YouTube footage, we extract 1.0M hand images from 20K videos. These extensive datasets provide diverse hand-object interactions across different views. We also prepare pre-training data with varying amount. "Exo-X" and "Ego-X" denote 100DOH and Ego4D datasets with X images selected randomly (*e.g.*, X = 50K, 100K, ..., 1M, 2M). "Ego&Exo-2M" shows our final set combining both datasets with full images (*i.e.*, 1.0M for each).

Fine-tuning datasets: We conduct fine-tuning experiments on three datasets with 3D hand pose ground truth in various data size and viewpoints: exocentric datasets from FreiHand [421] and DexYCB [49], and an egocentric dataset AssemblyHands [252]. FreiHand consists of 130.2K training frames and 3.9K test frames, with both green screen and real-world backgrounds. DexYCB contains 325.3K training images and 98.2K test images, focusing on natural hand-object interactions. AssemblyHands, the largest of the three, includes 704.0K training samples and 109.8K test samples, collected in object assembly scenarios. Following [316], we prepare 10% of the labeled FreiHand dataset, which is denoted as "FreiHand*", especially used for ablation studies. This allow us to assess the performance in a limited supervision setting.

Implementation details: For similar hands mining, we choose the PCA embedding size as $D = 14$. For the pre-training framework, we use ResNet-50 [131] as the encoder. Throughout the pre-training phase, all models are trained using LARS [386] with ADAM [163] optimizer, with the learning rate of 3.2e-3. Following [316], SimCLR employs scale and color jitter as image augmentation, while PeCLR and SiMHand utilize scale, rotation, translation, and color jitter. We use resized images with 128×128 as the input. We set the temperature parameter τ of contrastive learning as 0.5. We use 8 NVIDIA V100 GPUs with a batch size of 8192 for pre-training.

For fine-tuning, we initialize our model with the pre-trained encoder $E(\cdot)$ and

Method	Pre-training	FreiHand (Exo)		DexYCB (Exo)		AssemblyHands (Ego)	
		MPJPE ↓	*PCK-AUC* ↑	*MPJPE* ↓	*PCK-AUC* ↑	*MPJPE* ↓	*PCK-AUC* ↑
w/o pre-training	-	19.21	85.61	19.36	84.80	19.17	85.61
SimCLR	Exo-1M	19.30	85.36	20.13	83.75	20.01	84.21
	Ego-1M	19.36	85.09	20.22	83.50	20.32	83.85
	Ego&Exo-2M	20.07	84.32	21.09	82.25	21.24	82.29
PeCLR	Exo-1M	19.58	84.71	18.39	86.33	19.12	85.64
	Ego-1M	19.07	85.62	18.99	85.40	19.20	85.57
	Ego&Exo-2M	18.19	86.76	18.06	86.82	18.88	86.03
SiMHand (Ours)	Exo-1M	16.73	88.66	17.34	87.84	18.50	86.56
	Ego-1M	<u>16.15</u>	<u>89.48</u>	<u>16.99</u>	<u>88.34</u>	<u>18.26</u>	<u>86.95</u>
	Ego&Exo-2M	**15.79**	**90.04**	**16.71**	**88.86**	**18.23**	**86.90**

Table 5.1: **Comparison with the state of the art.** We show 3D hand pose estimation accuracy (MPJPE↓) on the FreiHand (Exo) [421], DexYCB (Exo) [49] and AssemblyHands (Ego) [252] . The best results are highlighted in **bold**, and the second-best results are <u>underlined</u>. SiMHand achieves the best results across various datasets.

then fine-tune with a 3D pose regressor on the labeled datasets. The 3D regressor involves 2D heatmap regression and 3D localization heads, similar to DetNet [412]. We use a single NVIDIA V100 GPU with a batch size of 128.

Evaluation: We use the following evaluation metrics: the mean per joint position error (MPJPE) in millimeters, which compares model predictions against ground-truth data, and the percentage of correct keypoints based on the area under the curve (PCK-AUC), which measures the proportion of predicted keypoints that fall within a specified distance (20mm to 50mm) from the ground truth with varying thresholds.

5.4.2 Main results

We compare our method with previous works for 3D hand pose estimation (Tab. 5.1). To make a fair comparison, we evaluate all pre-training datasets of the same size against previous methods.

Method	Pre-training size	FreiHand*	
		MPJPE ↓	*PCK-AUC* ↑
w/o pre-training	-	48.19	49.17
SimCLR	Ego-50K	53.94	42.54
PeCLR		47.42	49.85
SiMHand		**35.32**	**63.35**
SimCLR	Ego-100K	53.49	43.12
PeCLR		46.00	51.50
SiMHand		**31.06**	**68.66**
SimCLR	Ego-500K	49.91	47.61
PeCLR		43.18	54.15
SiMHand		**28.27**	**72.97**
SimCLR	Ego-1M	46.17	50.62
PeCLR		34.42	64.93
SiMHand		**23.68**	**79.62**

Table 5.2: **Comparison with different pre-training data sizes.** '*' indicates that we use a small amount of training data for fine-tuning to validate the effectiveness of the pre-trained model. Our method demonstrates a leading advantage across all pre-training data scales.

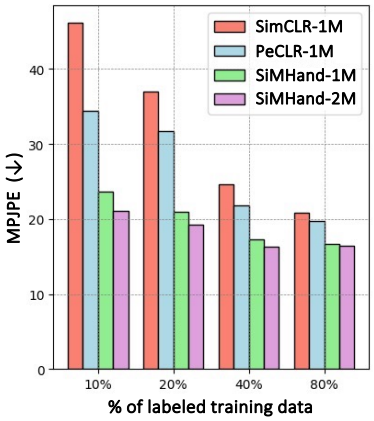


Figure 5.4: **Comparison with different data availability in fine-tuning on FreiHand.** Variations in the percentage of labeled data correspond to different subsets of the fine-tuning dataset, following the experimental design in [316].

Comparison to contrastive learning methods: We compare our pre-training method with previous methods [59, 316] in 3D hand pose estimation (Tab. 5.1). We observe that our method significantly outperforms SimCLR and PeCLR across various datasets under the equal pre-training data setups. When we compare our method against a randomly initialized model (w/o pre-training), SiMHand improves performance by 17.7% over the scratch baseline.

In more details, our approach achieves a 15.31% improvement over previous methods PeCLR with Ego-1M pre-training on the FreiHand. We observe that SimCLR shows limited performance compared to the random initialization. This suggests pre-training without geometric prior (*i.e.*, without geometric augmentation) does not always help hand pose estimation, requiring spatial keypoint regression. In contrast, our method demonstrates significant performance gain on larger datasets, with a 10.53% gain on DexYCB and a 4.90% improvement on Assem-

Method (Pre-training size)	Proposals		FreiHand*	
	Similar hands	**Adaptive weighting**	*MPJPE* ↓	*PCK-AUC* ↑
SimCLR	×	×	53.49	43.12
(Ego-100K)	×	✓	**52.58 (1.8% ↓)**	**44.70 (1.58% ↑)**
PeCLR	×	×	46.00	51.50
(Ego-100K)	×	✓	**44.61 (3.0% ↓)**	**53.37 (1.87% ↑)**
SiMHand	✓	×	31.06	68.66
(Ego-100K)	✓	✓	**28.84 (7.18% ↓)**	**71.07 (2.41% ↑)**

Table 5.3: **Ablation study of proposed modules.** We compare with and without our proposed modules in different methods. The experimental results demonstrate the generality of our method.

blyHands compared to PeCLR. These results confirm that our model consistently achieves superior performance across various fine-tuning datasets.

Furthermore, we pre-train all methods on the joint pre-training datasets (Ego&Exo-2M). Our approach further improves over the state-of-the-art method (PeCLR), achieving improvements of 13.19%, 7.4%, and 3.4% on the FreiHand, DexYCB, and AssemblyHands, respectively. Compared to the pre-training with 1M samples (Ego-1M), doubling the pretraining data with Ego&Exo-2M results in a 2.28% improvement on the FreiHand dataset. Notably, our method shows particular strength in effectively handling larger, more varied datasets. This robust performance demonstrates that our approach is highly effective and reliable for hand pose pre-training.

Ego & Exo view analysis: We evaluate how pre-training with egocentric views (Ego4D) and exocentric views (100DOH) affects the performance in datasets with their corresponding views, namely AssemblyHands for egocentric and FreiHand and DexYCB for exocentric views. Interestingly, matching pre-training viewpoints does not consistently enhance performance, indicating that the view gaps have limited effects. Instead, factors like dataset diversity and the characteristics of pre-training methods are more crucial in boosting performance. Combining the two datasets (the last row of Tab. 5.1) leads to the best performance in all three datasets, underscoring the potential of enriching data diversity with various camera views.

5.4.3 Ablation experiments

This section presents ablation studies on SiMHand, focusing on four aspects: 1) pre-training dataset size, 2) fine-tuning dataset size, 3) adaptive weighting, and 4) Top-K similar hands. First, we examine the size of the pre-training dataset using various methods, showing that our approach maintains superior performance across different sizes (Tab. 5.2). Second, inspired by [421], we explore fine-tuning dataset size, demonstrating significant gains even with limited data (Fig. 5.4). Furthermore, we also highlight the adaptive weighting design, which consistently outperforms comparison methods (Tab. 5.3). Finally, we conduct ablation analysis according to different levels of similarity in the assigned positive hand pairs. (Tab. 5.4).

Effect of pre-training data size: We study results with different sizes of pre-training data, namely 50K, 100K, 500K, and 1M in Tab. 5.2. The results demonstrate that SiMHand reliably outperforms the other methods across all settings, with improvement as the pre-training data size increases. With changes in the size of the pre-training data from 50K to 1M, SiMHand achieves a reduction in MPJPE from 35.32 to 23.68. The useful insights we can gather from this table include: 1) The SiMHand method holds a leading advantage across various pre-training size. 2) As the size of the pre-training dataset increases, the improvement for fine-tuning with limited labels is substantial.

Effect of fine-tuning data size: Fig. 5.4 illustrates the experiment under different proportions of labeled fine-tuning data, namely 10%, 20%, 40%, and 80% in FreiHand. Note that we denote methods with "-1M/2M" as those pre-trained on the Ego-1M and the Ego&Exo-2M sets, respectively. The results show that SiMHand-1M brings error reduction, achieving remarkably lower MPJPE scores with merely 10% of labeled data. SiMHand-1M delivers the best performance over different size of fine-tuning data, compared to SimCLR-1M and PeCLR-1M. SiMHand-2M further shows improvement over SiMHand-1M, while the gains become marginal as labeled data increase. From this analysis, we can draw two key conclusions: 1) The improvement resulting from an increase of pre-training data becomes less significant as the amount of fine-tuning data increases; 2) SiMHand maintains a strong advantage in scenarios with limited labeled data, particularly when larger pre-training data are used.

Method (Pre-training size)	Top-K	FreiHand* *MPJPE* ↓	*PCK-AUC* ↑
SiMHand (Ego-100K)	**Top-1**	**31.06**	**68.66**
	Top-2	31.46	67.89
	Top-5	31.85	67.20
	Top-10	31.87	67.18
	Top-50	31.53	67.59
	Top-100	31.54	67.70
	Top-500	32.61	66.76
	Top-1000	34.05	65.14
	Top-5000	35.34	62.79

Table 5.4: **Pre-training performance at different similarity ranks (Top-K).** It can be seen that as the similarity rank increases, the pre-training performance deteriorates.

Effect of adaptive weighting: We validate the proposed adaptive weighting and its generality when applied to the other methods in Tab. 5.3. On the Ego-100K pre-training set, the MPJPE scores after adaptive weighting decrease by 1.8% and 3.0% for SimCLR and PeCLR, respectively, while PCK-AUC increases by 1.58% and 1.87%. This indicates that the proposed weighting excels in its applicability to various pre-training methods. In our SiMHand method, applying adaptive weighting reduces MPJPE from 31.06 to 28.84, a 7.18% decrease, while PCK-AUC improves from 68.66 to 71.07, a 2.41% increase. We find the effectiveness of the proposed weighting when combined with the mined similar hands.

Learning from Top-K similar hands: We test pre-training with different similarity levels of positive samples in Tab. 5.4. As illustrated in Fig. 5.2, we can sample similar pairs according to the distance ranking (*e.g.*, K = 1, 2, ..., 5000), where Top-1 is used to produce our final results. The performance trend is initially subtle and somewhat fluctuating (Top-1∼100) but becomes increasingly pronounced after Top-100. This indicates that as the similarity between positive samples in-

creases, the global trend decreases accordingly. Notably, using Top-5000 similar hand samples as positive samples decreases the MPJPE by 13.78% compared to Top-1. This study provides two insights: 1) Similar samples with subtle noisiness (*e.g.*, 1∼100) exhibit minimal variation in performance, indicating that slight differences in similarity within this range do not significantly impact the pre-training outcome. This suggests that the model is robust to minor variations when the positive samples are highly similar. 2) The results support the validity of using Top-1 positive samples to produce final results, as they consistently exhibit the best performance. This highlights the importance of selecting the most similar samples in contrastive learning.

5.4.4 Visualization

In this section, we compare the fine-tuning results of various pre-training methods through detailed visualizations on different datasets (Fig. 5.5). The pre-training model is trained on the Ego&Exo-2M dataset and fine-tuned on the FreiHands [421] and DexYCB [49] datasets, respectively.

From the left four columns of Fig. 5.5, the visualization results show that SiMHand performs better in pose estimation, with results closer to the ground truth, compared to the other methods in FreiHands [421] dataset. In particular, SiMHand outperforms the other methods in challenging environments, such as those with varying lighting conditions, by better capturing hand poses. These visual outputs highlight its robustness across various scenarios, solidifying its potential for real-world applications.

As shown in the right four columns of Fig. 5.5, we highlight the occluded regions in the original images of DexYCB [49] dataset using red circles. The results show that SiMHand is more effective in tackling occlusion problems. Our pre-training method effectively addresses partially occluded images by utilizing similar, though not identical, hand images, where the occluded parts in the query image may be visible in the corresponding similar hand image, and vice versa.

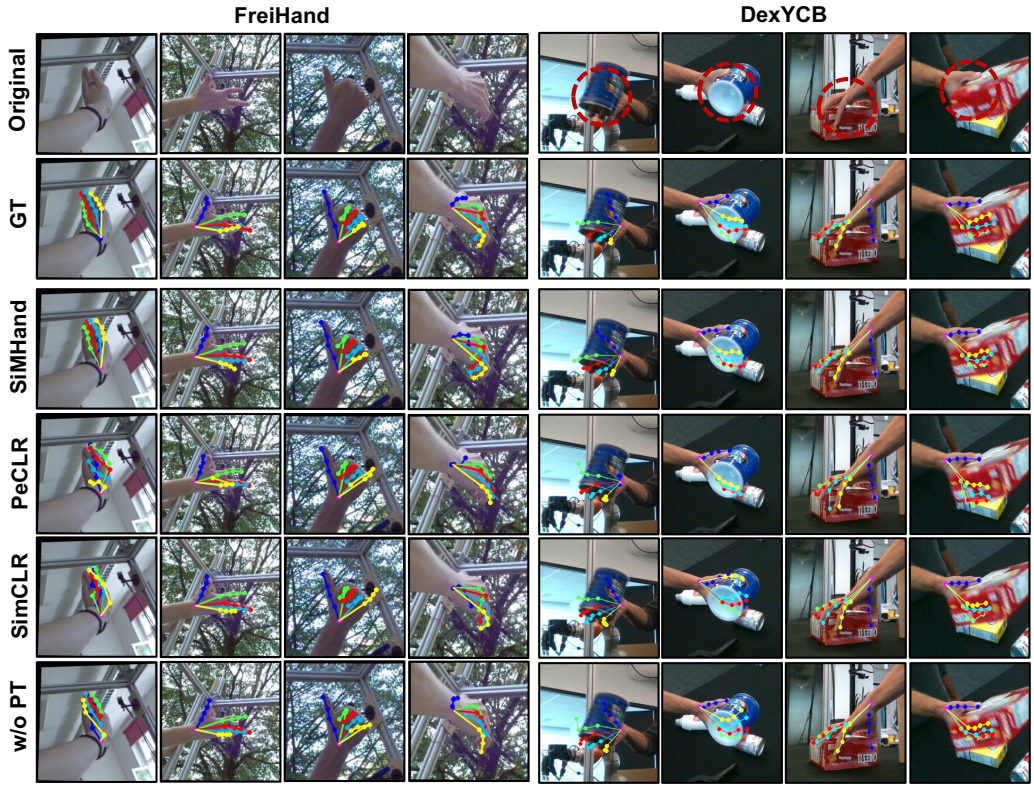


Figure 5.5: **Visualization of FreiHand [421] and DexYCB [49].** The first four columns on the left display the results for FreiHand, while the last four columns on the right show the results for DexYCB (GT: Ground Truth; PT: Pre-training). It can be observed that SiMHand pre-training method achieves better results.

Figure 5.6: **Overview of data preprocessing and similar hands mining.** This image illustrates a three-step process for SiMHand pre-training using datasets from Ego4D and 100DOH. **Step 1** involves preprocessing the datasets to extract relevant frames. **Step 2** employs a hand detector to crop hand regions from these frames, creating a diverse pool of hand images in the wild. **Step 3** calculates similarity and ranks the images using a pose estimator and PCA, producing a sorted list of hand poses, from the most similar to the least similar to a given anchor pose.

5.5 Additional Results and Details

5.5.1 Construction of large-scale in-the-wild hand database

This section presents our method for constructing a large-scale hand image dataset by extracting and processing hand images from various video datasets. We outline key preprocessing steps, including *1) preprocessing*, *2) hand region detection*, and *3) similarity calculation & ranking*.

Preprocessing: We prepare two large-scale video datasets: Ego4D, containing 8K frames, and 100DOH, with 23K frames, both sampled at 1 *fps*. As shown in Fig. 5.6, first-person and third-person hand images exhibit significant differences.

Hand region detection: After extracting frames from Ego4D and 100DOH, we use a lightweight, fixed-weight network to detect hand regions via bounding boxes. Specifically, we adopt the method from [305] and store all detected bounding boxes in sequence. This step constructs a large-scale hand image dataset as [328].

Similarity calculation & ranking: Once the hand image dataset is built, we use a lightweight, fixed-weight network to extract raw keypoints for each sample via MediaPipe [209]. To reduce noise, we apply PCA as described in Sec. 5.3.1. We then compute similarity scores for a given query image I using Eq. 5.1 and rank the remaining samples accordingly. This process yields a large-scale set of in-the-wild hand images with similar characteristics. For instance, in Ego4D, given a query sample I, we retrieve all similar hand images and construct a ranked sequence, referred to as "Top-K". The Top-1 image in this sequence serves as the positive sample I^+ for contrastive learning, enhancing the effectiveness of SiMHand pre-training. As shown in Tab. 5.4, our experiments validate that selecting Top-1 as the positive sample I^+ is the optimal strategy.

5.5.2 Finetuning for 3D hand pose estimation

In the fine-tuning stage, we discard the projection head and fine-tuning only the encoders. We load the pre-training model weights into a heatmap-based 3D hand pose estimationand predition method: DetNet[412]. To train DetNet, we utilize a comprehensive loss function designed to optimize both 2D pose estimation and 3D spatial localization. The loss function is defined as:

$$\mathcal{L}_{\text{heat}} + \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{delta}} + \mathcal{L}_{\text{reg}} \tag{5.4}$$

where $\mathcal{L}_{\text{heat}}$ ensures that the predicted heatmaps H align closely with the ground truth heatmaps H^{GT}, $\mathcal{L}_{\text{loc}}$ and $\mathcal{L}_{\text{delta}}$ measure the discrepancies between the predicted location maps L and delta maps D and their corresponding ground truth L^{GT} and D^{GT}, with H^{GT} weighting these discrepancies to focus on the maxima of the heatmaps. Additionally, $\mathcal{L}_{\text{reg}}$ is an $L2$ regularization term to prevent overfitting. Note that after passing through the encoder, we made simple adjustments to the model, applying some upsampling to the features to fit the input.

This multi-task learning framework enables the network to simultaneously learn pose features from 2D images and spatial information from 3D data, enhancing the accuracy and robustness of detection in real-world applications. For more details on fine-tuning, please refer to the [412].

Method	Pre-training size	FreiHand*	
		MPJPE ↓	*PCK-AUC* ↑
PeCLR	Ego-50K	47.42	49.85
TempCLR		45.17	52.40
SiMHand		**35.32**	**63.35**
PeCLR	Ego-100K	46.00	51.50
TempCLR		44.54	53.28
SiMHand		**31.06**	**68.66**

Table 5.5: **Comparison with the TempCLR method.** "*" indicates that we use a small amount of training data for fine-tuning to validate the effectiveness of the pre-trained model. TempCLR outperforms PeCLR by a modest margin, whereas SiMHand achieves a significant performance improvement over TempCLR.

5.5.3 Comparison with TempCLR method

We conduct an experimental comparison with the TempCLR [418] method. TempCLR proposes a pre-training framework for 3D hand reconstruction using time-coherent contrastive learning and demonstrates better performance compared to PeCLR [316]. Although TempCLR primarily focuses on reconstruction tasks, the parametric model it uses can also output 3D pose results, making it valuable to further compare our method with TempCLR.

However, TempCLR has certain limitations in data collection and the effectiveness of contrastive learning. First, TempCLR treats hands from adjacent frames as positive samples during training. In dynamic egocentric videos, hand occlusions or detection failures often lead to missed hand crops in neighboring frames. In addition, images from adjacent frames typically lack background diversity, limiting the contribution of positive sample pairs formed from neighboring frames in contrastive learning.

In contrast to TempCLR, our method, SiMHand, significantly improves performance. SiMHand leverages similar hand images, which provide richer diversity in features, including various types of hand-object interactions, diverse backgrounds, and varying appearances. These features allow SiMHand to effectively

Method	Backbone	DexYCB *MPJPE* ↓
[366]	ResNet50	25.57
[317]	ResNet50	22.71
[317]	HRNet32	22.26
[335]	ResNet18	21.22
[412]	ResNet50	19.36
SiMHand	ResNet50	**16.71**

Table 5.6: **Comparison of 3D hand pose estimation methods on DexYCB [49].**

increase the diversity of positive samples in contrastive learning, resulting in superior pre-training performance.

We further validate our approach on two different size of pre-training data, consisting of 50K and 100K hand images from the Ego4D dataset [111]. Tab. 5.5 shows the significant progress made by SiMHand compared to TempCLR and PeCLR.

From the experimental results, TempCLR demonstrates better performance than PeCLR, which matches the conclusion of the original paper. However, SiMHand provides more valuable positive samples for contrastive learning, leading to better results during the fine-tuning phase of 3D hand pose estimation tasks.

5.5.4 Comparison with the other 3D hand pose estimation methods

To better assess the value of this work and its position within the broader context, we have included comparisons with other related works in the field of 3D hand pose estimation in this section.

As shown in Tab. 5.6 and 5.7, the comparative results on DexYCB [49] and AssemblyHands [252] datasets further validate the superiority of our approach across multiple standard datasets, demonstrating the effectiveness of our pretraining strategy and its broad potential for real-world applications.

Method	Backbone	AssemblyHands *MPJPE* ↓
[119]	ResNet50	32.91
[252]	ResNet50	21.92
[412]	ResNet50	19.17
SiMHand	ResNet50	**18.23**

Table 5.7: **Comparison of 3D hand pose estimation methods on AssemblyHands [252].**

5.5.5 Visualization on AssemblyHands

Fig. 5.7 shows the visualization results of hand pose estimation on another dataset, AssemblyHands [252]. We highlight instances of hand-object occlusion in the data using red circles. As observed with DexYCB [49], SiMHand pre-trained model demonstrates superior performance in handling occlusion during the fine-tuning stage compared to the other pre-training methods, showcasing stronger robustness.

5.5.6 Visualization of similar hands

We present the visualization of Top-K similar hand images used to create positive pairs. As shown in Fig. 5.8, we visualize a set of Top-K similar hand images. The figure displays the query image alongside its corresponding similar hand sequence (Top-K). At the top of Fig. 5.8, a timeline indicates that the images are deliberately sampled from consecutive frames of the same video.

From these visualizations, we derive three key insights: 1) Using adjacent frames from the same video as positive samples in pre-training lacks diversity, as substantial variations may still exist between samples. 2) As the ranking increases, the similarity between hand images decreases significantly, leading to greater differences that may result in inaccurate feature representations during pre-training. 3) Therefore, selecting the Top-1 image is a proper design to assign diverse yet similar positive samples for the query images.

Figure 5.7: **Visualization of Hand Pose Estimation Results on AssemblyHands.** AssemblyHands [252] is a hand pose dataset captured from a first-person perspective during toy assembly. It can be observed that SiMHand pre-training method achieves better results (GT: Ground Truth; PT: Pre-training).

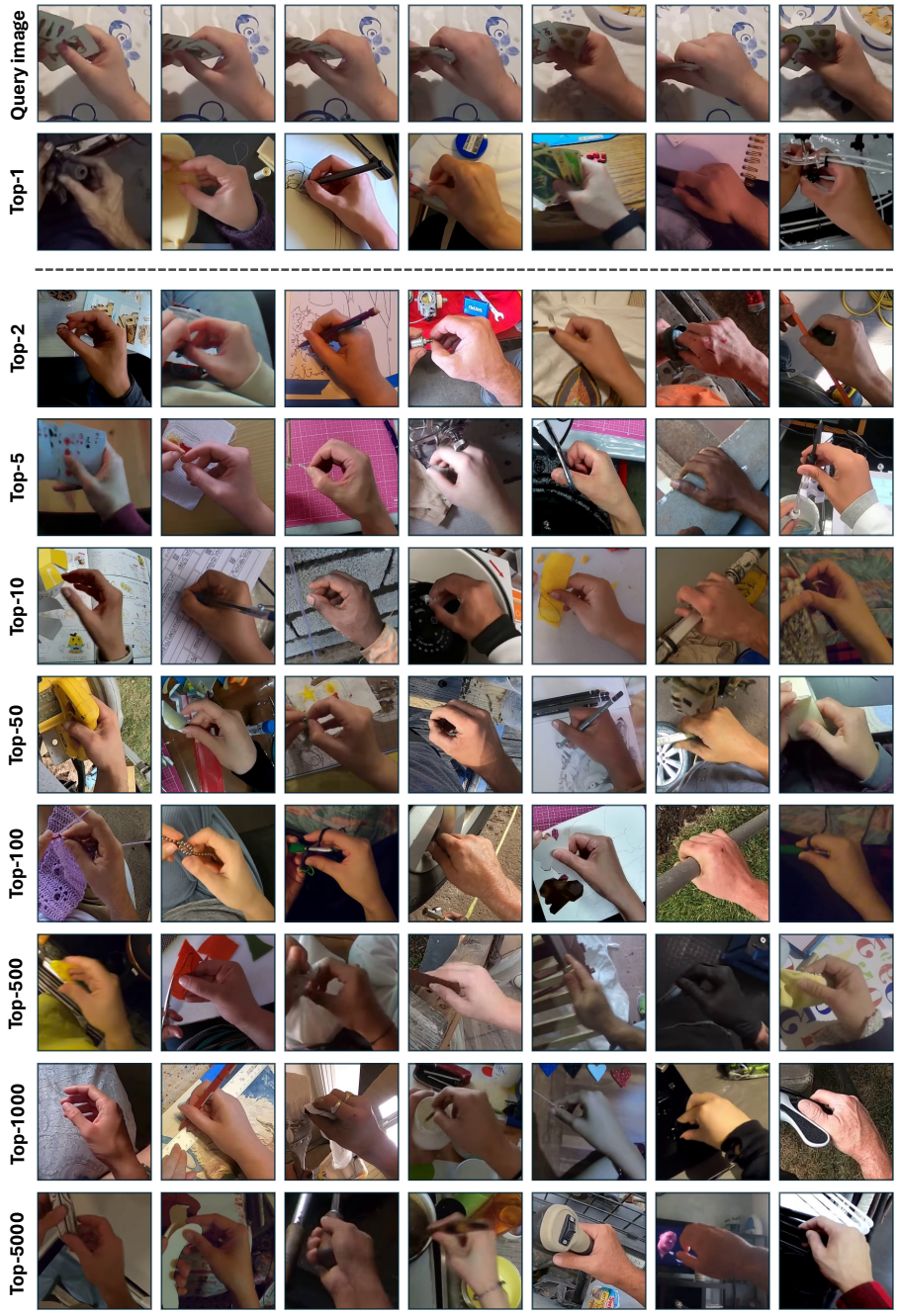


Figure 5.8: **Visualization of similar hand samples in Top-K.** As the ranking increases, the differences between hand samples become more pronounced.

5.6 Conclusion

We introduce SiMHand, a contrastive learning framework for pre-training 3D hand pose estimators by mining similar hand pairs from large-scale in-the-wild images. Our approach leverages similar hand pairs from diverse videos, significantly enhancing the information gained during pre-training compared with existing methods. Experiments show that our pre-training method achieves competitive performance in 3D hand pose estimation across multiple datasets, outperforming previous pre-training approaches and demonstrating the benefits of large-scale pre-training with in-the-wild images. We hope this work can lay a foundation for future research on pre-training of 3D hand pose estimation.

Chapter 6

Domain Adaptive Hand State Estimation in the Wild

In this chapter, we aim to improve the performance of regressing hand keypoints and segmenting pixel-level hand masks under new imaging conditions (*e.g.*, outdoors) when we only have labeled images taken under very different conditions (*e.g.*, indoors). In the real world, it is important that the model trained for both tasks works under various imaging conditions. However, their variation covered by existing labeled hand datasets is limited. Thus, it is necessary to adapt the model trained on the labeled images (source) to unlabeled images (target) with unseen imaging conditions. While self-training domain adaptation methods (*i.e.*, learning from the unlabeled target images in a self-supervised manner) have been developed for both tasks, their training may degrade performance when the predictions on the target images are noisy. To avoid this, it is crucial to assign a low importance (confidence) weight to the noisy predictions during self-training. In this paper, we propose to utilize the divergence of two predictions to estimate the confidence of the target image for both tasks. These predictions are given from two separate networks, and their divergence helps identify the noisy predictions. To integrate our proposed confidence estimation into self-training, we propose a teacher-student framework where the two networks (teachers) provide supervision to a network (student) for self-training, and the teachers are learned from the student by knowledge distillation. Our experiments show its superiority over state-of-the-art methods in adaptation settings with different lighting, grasping objects,

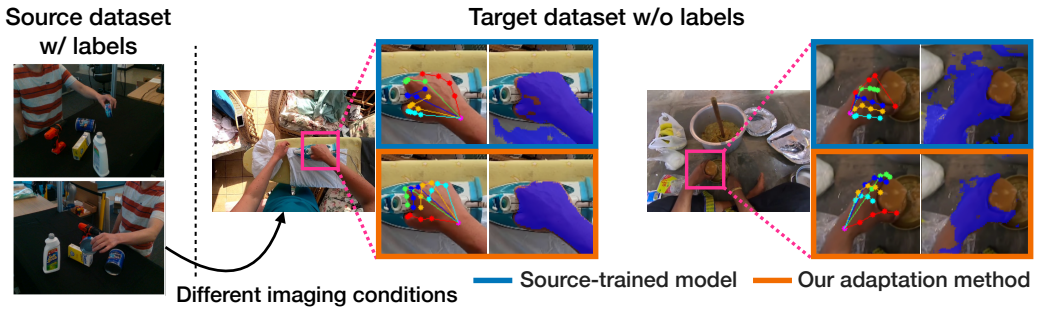


Figure 6.1: **Overview**. We aim to adapt the model of localizing hand keypoints and pixel-level hand masks to new imaging conditions without annotation.

backgrounds, and camera viewpoints. Our method improves by 4% the multi-task score on HO3D compared to the latest adversarial adaptation method. We also validate our method on Ego4D, egocentric videos with rapid changes in imaging conditions outdoors.

6.1 Introduction

In the real world, hand keypoint regression and hand segmentation are considered important to work under broad imaging conditions for various computer vision applications, such as egocentric video understanding [77, 111], hand-object interaction analysis [45, 97], AR/VR [190, 362], and assistive technology [176, 191]. For building models for both tasks, several labeled hand datasets have been proposed in laboratory settings, such as multi-camera studios [49, 155, 230, 421] and attaching sensors to hands [103, 107, 392]. However, their imaging conditions do not adequately cover real-world imaging conditions [256], consisting of various lighting, hand-held objects, backgrounds, and camera viewpoints. In addition, the annotation of keypoints and pixel-level masks are not always available in real-world environments because they are labor-intensive to acquire. As shown in Fig. 6.1, when localizing hand keypoints and pixels in real-world egocentric videos [111] (*e.g.*, outdoors), we may only have access to a hand dataset [49] taken under completely different imaging conditions (*e.g.*, indoors). Given these limitations, we need methods that can robustly adapt the models trained on the available labeled images (source) to unlabeled images (target) with new imaging

conditions.

To enable such adaptation, the approach of self-training domain adaptation has been developed for both tasks. This approach aims to learn unlabeled target images by optimizing a self-supervised task, which exhibits effectiveness in various domain adaptation tasks [37, 56, 79, 347, 409]. For keypoint estimation, consistency training, a method that regularizes keypoint predictions to be consistent under geometric transformations, has been proposed [343, 377, 409]. As for hand segmentation, prior studies use pseudo-labeling [37, 256], which produces hard labels by thresholding a predicted class probability for updating a network. However, these self-training methods for both tasks perform well only when the predictions are reasonably correct. When the predictions become noisy due to the gap in imaging conditions, the trained network will cause over-fitting to the noisy predictions, resulting in poor performance in the target domain.

To avoid this, it is crucial to assign a low importance (confidence) weight to the loss of self-training with noisy predictions. This confidence weighting can mitigate the distractions from the noisy predictions. To this end, we propose self-training domain adaptation with confidence estimation for hand keypoint regression and hand segmentation. Our proposed method consists of (i) confidence estimation based on the divergence of two networks' predictions and (ii) an update rule that integrates a training network for self-training and the two networks for confidence estimation.

To (i) estimate confidence, we utilize the predictions of two different networks. While class probability can be used as the confidence in classification tasks, it is not trivial to obtain such a measure in keypoint regression. Thus, we newly focus on the divergence of the two networks' predictions for each target image. We design their networks to have an identical architecture but have different learning parameters. We observe that when the divergence measure is high, the predictions of both networks are noisy and should be avoided in self-training.

To (ii) integrate the estimated confidence into self-training, inspired by the single-teacher-single-student update [268, 329], we develop mutual training with self-training based on consistency training for a training network (student) and distillation-based update for the two networks (teachers). For training the student network, we build a unified self-training framework that can work favorably for the two tasks. Motivated by supervised or weakly-supervised learning for jointly

estimating both tasks [63, 108, 241, 357, 396], we expect that jointly adapting both tasks will allow one task to provide useful cues to the other task even in the unlabeled target domain. Specifically, we enforce the student network to generate consistent predictions for both tasks under geometric augmentation. We weight the loss of the consistency training using the confidence estimated from the divergence of the teachers' predictions. This can reduce the weight of the noisy predictions during the consistency training. To learn the two teacher networks differently, we train the teachers independently from different mini-batches by knowledge distillation, which matches the teacher-student predictions in the output level. This framework enables the teachers to update more carefully than the student and prevent over-fitting to the noisy predictions. Such stable teachers provide reliable confidence estimation for the student's training.

In our experiments, we validate our proposed method in adaptation settings where lighting, grasping objects, backgrounds, camera viewpoints, etc., vary between labeled source images and unlabeled target images. We use a large-scale hand dataset captured in a multi-camera system [49] as the source dataset (see Fig. 6.1). For the target dataset, we use HO3D [116] with different environments, HanCo [419] with multiple viewpoints and diverse backgrounds, and FPHA [103] with a novel first-person camera viewpoint. We also apply our method to in-the-wild egocentric video Ego4D [111] (see Fig. 6.1), including diverse indoor and outdoor activities worldwide. Our method improves the average score of the two tasks by 14.4%, 14.9%, and 18.0% on HO3D, HanCo, and FPHA, respectively, compared to a unadapted baseline. Our method further exhibits distinct improvements compared to the latest adversarial adaption method [151] and consistency training baselines with uncertainty estimation [37], confident instance selection [256], and the teacher-student scheme [329]. We finally confirm that our method also performs qualitatively well on the Ego4D videos.

Our contributions are summarized as follows:

- We propose a novel confidence estimation method based on the divergence of the predictions from two teacher networks for self-training domain adaptation of hand keypoint regression and hand segmentation.
- To integrate our proposed confidence estimation into self-training, we propose mutual training using knowledge distillation with a student network for self-training and two teacher networks for confidence estimation.

- Our proposed framework outperforms state-of-the-art methods under three adaptation settings across different imaging conditions. It also shows improved qualitative performance on in-the-wild egocentric videos.

6.2 Related Work

Hand keypoint regression: Hand keypoint regression is the task of regressing the positions of hand joint keypoints from a cropped hand image. 2D hand keypoint regression is trained by optimizing keypoint heatmaps [242, 358, 420] or directly predicting keypoint coordinates [298]. The 2D keypoints are informative for estimating 3D hand poses [31, 233, 309, 380]. To build an accurate keypoint regressor, collecting massive hand keypoint annotations is required but laborious. While early works annotate the keypoints manually from a single view [235, 274, 320], recent studies have collected the annotation more densely and efficiently using synthetic hand models [127, 233, 235, 420], hand sensors [103, 107, 324, 392], or multi-camera setups [33, 49, 116, 155, 208, 230, 421]. However, these methods suffer the gap in imaging conditions with real-world images in deployment [256]. For instance, the synthetic hand models and hand sensors induce different lighting conditions from actual human hands. The multi-camera setup lacks a variety of lighting, grasping objects, and backgrounds. To tackle these problems, domain adaptation is a promising solution that can transfer the knowledge of the network trained on source data to unlabeled target data. Jiang *et al*. proposed an adversarial domain adaptation for human and hand keypoint regression, optimizing the discrepancy between regressors [151]. Additionally, self-training adaptation methods have been studied in the keypoint regression of animals [43], humans [343], and objects [409]. Unlike these prior works, we incorporate confidence estimation into a self-training method based on consistency training for keypoint regression.
Hand segmentation: Hand segmentation is the task of segmenting pixel-level hand masks in a given image. CNN-based segmentation networks [24, 162, 341] are popularly used. The task can be jointly trained with hand keypoint regression because detecting hand regions guides to improve keypoint localization [63, 108, 241, 357, 396]. Since hand mask annotation is laborious as hand keypoint regression, a few domain adaptation methods with pseudo-labeling have been explored [37, 256]. To reduce the effect of highly noisy pseudo-labels in the target

domain, Cai *et al*. incorporate the uncertainty of pseudo-labels in model adaptation [37], and Ohkawa *et al*. select confident pseudo-labels by the overlap of two predicted hand masks [256]. Unlike [37], we estimate the target confidence using two networks. Instead of using the estimated confidence for instance selection [256], we assign the confidence to weight the loss of consistency training.

Domain adaptation via self-training: Domain adaptation aims to learn unlabeled target data in a self-supervised learning manner. This approach can be divided into three categories. (i) Pseudo-labeling [37, 56, 256, 296, 422] learns unlabeled data with hard labels assigned by confidence thresholding from the output of a network. (ii) Entropy minimization [205, 270, 347] regularizes the conditional entropy of unlabeled data and increases the confidence of class probability. (iii) Consistency regularization [52, 96, 365] enforces regularization so that the prediction on unlabeled data is invariant under data perturbation. We choose to leverage this consistency-based method for our task because it works for various tasks [202, 217, 253] and the first two approaches cannot be directly applied. Similar to our work, Yang *et al*. [377] enforce the consistency for two different views and modalities in hand keypoint regression. Mean teacher [329] provides teacher-student training with consistency regularization, which regularizes a teacher network by a student's weights and avoids over-fitting to incorrect predictions. Unlike [377], we propose to integrate confidence estimation into the consistency training and adopt the teacher-student scheme with two networks. To encourage the two networks to have different representations, we propose a distillation-based update rule instead of updating the teacher with the exponential moving average [329].

6.3 Proposed Method

In this section, we present our proposed self-training domain adaptation with confidence estimation for adapting hand keypoint regression and hand segmentation. We first present our problem formulation and network initialization with supervised learning from source data. We then introduce our proposed modules: (1) geometric augmentation consistency, (2) confidence weighting by using two networks, and (3) teacher-student update via knowledge distillation. As shown in Fig. 6.2, our adaptation is done with two different networks (teachers) for confi-

dence estimation and another network (student) for self-training of both tasks.

Problem formulation: Given labeled images from one source domain and unlabeled images from another target domain, we aim to jointly estimate hand keypoint coordinates and pixel-level hand masks on the target domain. We have a source image $\boldsymbol{x}_{\mathrm{s}}$ drawn from a set $X_{\mathrm{s}} \subset \mathbb{R}^{H \times W \times 3}$, its corresponding labels $(\boldsymbol{y}_{\mathrm{s}}^{\mathrm{p}}, \boldsymbol{y}_{\mathrm{s}}^{\mathrm{m}})$, and a target image $\boldsymbol{x}_{\mathrm{t}}$ drawn from a set $X_{\mathrm{t}} \subset \mathbb{R}^{H \times W \times 3}$. The pose label $\boldsymbol{y}_{\mathrm{s}}^{\mathrm{p}}$ consists of the 2D keypoint coordinates of 21 hand joints obtained from a set $Y_{\mathrm{s}}^{\mathrm{p}} \subset \mathbb{R}^{21 \times 2}$, while the mask label $\boldsymbol{y}_{\mathrm{s}}^{\mathrm{m}}$ denotes a binary mask obtained from $Y_{\mathrm{s}}^{\mathrm{m}} \subset (0,1)^{H \times W}$. A network parameterized by $\boldsymbol{\theta}$ learns the mappings $f^k(x;\boldsymbol{\theta}) : X \rightarrow Y^k$ where $k \in \{\mathrm{p}, \mathrm{m}\}$ represents the indicator for both tasks.

Initialization with supervised learning: To initialize networks used in our adaptation, we train the network f on the labeled source data following multi-task learning. Given the labeled dataset $(X_{\mathrm{s}}, Y_{\mathrm{s}})$ and the network $\boldsymbol{\theta}$, a supervised loss function is defined as

$$\mathcal{L}_{\mathrm{task}}\left(\boldsymbol{\theta}, X_{\mathrm{s}}, Y_{\mathrm{s}}\right) = \sum_{k} \lambda^k \mathbb{E}_{(\boldsymbol{x}_{\mathrm{s}}, \boldsymbol{y}_{\mathrm{s}}^k) \sim (X_{\mathrm{s}}, Y_{\mathrm{s}}^k)} \left[\mathcal{L}^k(\boldsymbol{p}_s^k, \boldsymbol{y}_{\mathrm{s}}^k)\right], \tag{6.1}$$

where $Y_{\mathrm{s}} = \{Y_{\mathrm{s}}^{\mathrm{p}}, Y_{\mathrm{s}}^{\mathrm{m}}\}$ and $\boldsymbol{p}_{\mathrm{s}}^k = f^k(\boldsymbol{x}_{\mathrm{s}};\boldsymbol{\theta})$. $\mathcal{L}^k(\cdot,\cdot) : Y^k \times Y^k \rightarrow \mathbb{R}^+$ is a loss function in each task and λ^k is a hyperparameter to balance the two tasks. We use a smooth L1 loss [145, 286] as $\mathcal{L}^{\mathrm{p}}$ and a binary cross-entropy loss as $\mathcal{L}^{\mathrm{m}}$.

6.3.1 Geometric augmentation consistency

Inspired by semi-supervised learning using hand keypoint consistency [377], we advance a unified training with consistency for both hand keypoint regression and hand segmentation. We expect that joint adaption of both tasks will allow one task to provide useful cues to the other task in consistency training, as studied in supervised or weakly-supervised learning setups [63, 108, 241, 357, 396]. We design consistency training by predicting the location of hand keypoints and hand pixels in a given geometrically transformed image, including rotation and transition. This consistency under geometric augmentation encourages the network to learn against positional bias in the target domain, which helps capture the hand structure related to poses and regions. Specifically, given a paired augmentation function $(T_{\mathrm{x}}, T_{\mathrm{y}}^k) \sim \mathcal{T}$ for an image and an label, we generate the prediction on the target images $\boldsymbol{p}_{\mathrm{t}}^k = f^k\left(\boldsymbol{x}_{\mathrm{t}};\boldsymbol{\theta}\right)$ and the augmented target images $\boldsymbol{p}_{\mathrm{t,aug}}^k = f^k\left(T_{\mathrm{x}}(\boldsymbol{x}_{\mathrm{t}});\boldsymbol{\theta}\right)$.

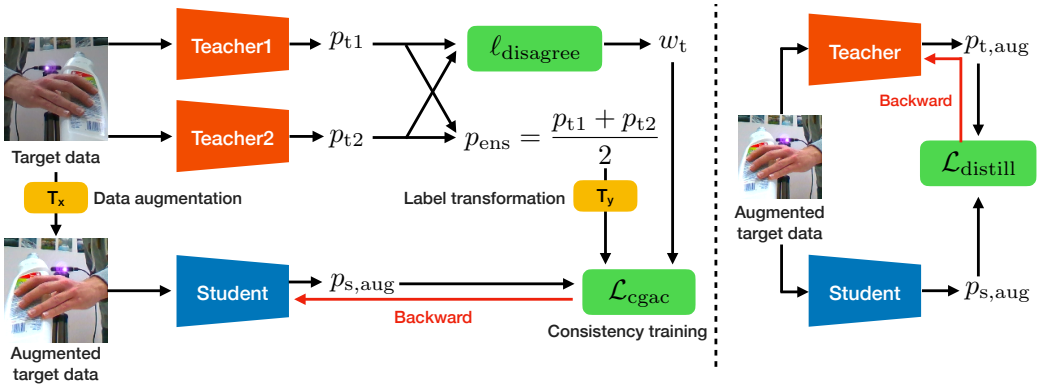


Figure 6.2: **Method overview**. **Left**: Student training with confidence-aware geometric augmentation consistency. The student learns from the consistency between its prediction and the two teachers' predictions. The training is weighted by the target confidence computed by the divergence of both teachers. **Right**: Teacher training with knowledge distillation. Each teacher independently learns to match the student's predictions. The task index k is omitted for simplicity.

We define the loss function of geometric augmentation consistency (GAC) $\mathcal{L}_{\text{gac}}$ between $\boldsymbol{p}^k_{\text{t,aug}}$ and $T^k_{\text{y}}(\boldsymbol{p}_{\text{t}})$ as

$$\mathcal{L}_{\text{gac}}\left(\boldsymbol{\theta}, X_t, \mathcal{T}\right) = \mathbb{E}_{\boldsymbol{x}_{\text{t}},(T_{\text{x}},T^{\text{p}}_{\text{y}},T^{\text{m}}_{\text{y}})}\left[\sum_{k\in\{\text{p,m}\}} \tilde{\lambda}^k \tilde{\mathcal{L}}^k\left(\boldsymbol{p}^k_{\text{t,aug}}, T^k_{\text{y}}(\boldsymbol{p}^k_{\text{t}})\right)\right]. \tag{6.2}$$

To correct the augmented prediction $\boldsymbol{p}^k_{\text{t,aug}}$ by $T^k_{\text{y}}(\boldsymbol{p}_{\text{t}})$, we stop the gradient update for $\boldsymbol{p}^k_{\text{t}}$, which can be viewed as the supervision to $\boldsymbol{p}^k_{\text{t,aug}}$. We use the smooth L1 loss (see Eq. (6.1)) as $\tilde{\mathcal{L}}^{\text{p}}$ and a mean squared error as $\tilde{\mathcal{L}}^{\text{m}}$. We introduce $\tilde{\lambda}^k$ as a hyperparameter to control the balance of the two tasks. The augmentation set $\mathcal{T}$ contains the geometric augmentation and photometric augmentation, such as color jitter and blurring. We set $T_{\text{y}}(\cdot)$ to align geometric information to the augmented input $T_{\text{x}}(\boldsymbol{x}_{\text{t}})$. For example, we apply rotation $T_{\text{y}}(\cdot)$ to the outputs $\boldsymbol{p}^k_{\text{t}}$ with the same degree of rotation $T_{\text{x}}(\cdot)$ to the input $\boldsymbol{x}_{\text{t}}$.

6.3.2 Confidence estimation by two separate networks

Since the target predictions are not always reliable, we aim to incorporate the estimated confidence weight for each target instance into the consistency training.

In Eq. (6.2), the generated outputs $\boldsymbol{p}_{\mathrm{t}}^{k}$ that is the supervision to $\boldsymbol{p}_{\mathrm{t,aug}}^{k}$ may be unstable and noisy due to the domain gap between source and target domains. Due to that, the network trained with the consistency readily overfits to the incorrect supervision $\boldsymbol{p}_{\mathrm{t}}^{k}$, which is known as confirmation bias [10, 329]. To reduce the bias, it is crucial to assign a low importance (confidence) weight to the consistency training with the incorrect supervision. This enables the network to learn primarily from reliable supervision while avoiding being biased to such erroneous predictions. In classification tasks, predicted class probability can serve as the confidence, while these measures are not trivially defined and available in regression tasks. To estimate the confidence of keypoint predictions, Yang *et al.* [377] measure the confidence of 3D hand keypoints by the distance to the fitted 3D hand template, but the hand template fitting is an ill-posed problem for 2D hands and is not applicable to hand segmentation. Dropout [37, 38, 101] is a generic way of estimating uncertainty (confidence), calculated by the variance of multiple stochastic forwards. However, the estimated confidence is biased to the current state of the training network because the training and confidence estimation are done by a single network. When the training network works poorly, the confidence estimation becomes readily unreliable.

To perform reliable confidence estimation for both tasks, we propose a confidence measure by computing the divergence of two predictions. Specifically, we introduce two networks (*a.k.a.*, teachers) for the confidence estimation and the estimated confidence is used to train another network (*a.k.a.*, student) for the consistency training. The architecture of the teachers is identical, yet they have different learning parameters. We observe that when the divergence of the two predictions from the teachers for a target instance is high, the predictions of both networks become unstable. In contrast, a lower divergence indicates that the two teacher networks predict stably and agree on their predictions. Thus, we use the divergence for representing the target confidence. Given the teachers $\boldsymbol{\theta}^{\mathrm{tch1}}, \boldsymbol{\theta}^{\mathrm{tch2}}$, we define a disagreement measure ℓ_{disagree} to compute the divergence as

$$\ell_{\mathrm{disagree}}\left(\boldsymbol{\theta}^{\mathrm{tch1}}, \boldsymbol{\theta}^{\mathrm{tch2}}, \boldsymbol{x}_{\mathrm{t}}\right) = \sum_{k \in \{\mathrm{p,m}\}} \tilde{\lambda}^{k} \tilde{\mathcal{L}}^{k}(\boldsymbol{p}_{\mathrm{t1}}^{k}, \boldsymbol{p}_{\mathrm{t2}}^{k}), \tag{6.3}$$

where $\boldsymbol{p}_{\mathrm{t1}}^{k} = f^{k}(\boldsymbol{x}_{\mathrm{t}}; \boldsymbol{\theta}^{\mathrm{tch1}})$ and $\boldsymbol{p}_{\mathrm{t2}}^{k} = f^{k}(\boldsymbol{x}_{\mathrm{t}}; \boldsymbol{\theta}^{\mathrm{tch2}})$.

As a proof of concept, we visualize the correlation between the disagreement measure and a validation score averaged over evaluation metrics of the two tasks

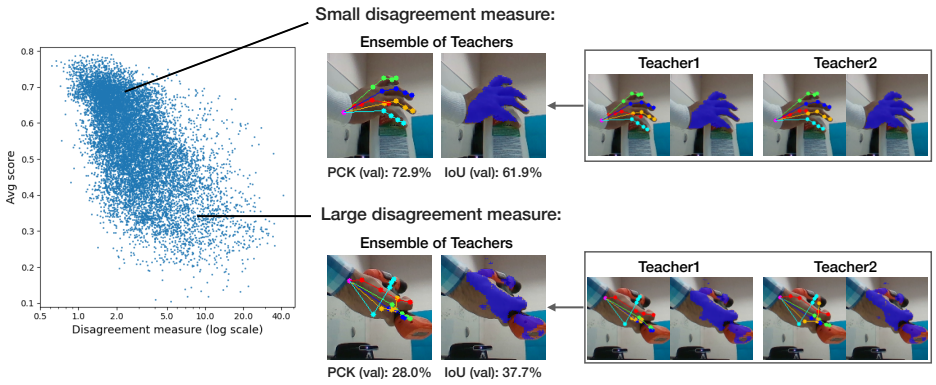


Figure 6.3: **The correlation between a disagreement measure and task scores.** Target instances with smaller disagreement values between the two teacher networks tend to have higher task scores.

(PCK and IoU) in Fig. 6.3. We compute the score between the ensemble of the teachers' predictions $\boldsymbol{p}^k_{\mathrm{ens}} = \left(\boldsymbol{p}^k_{\mathrm{t1}} + \boldsymbol{p}^k_{\mathrm{t2}}\right)/2$ and its ground truth in the validation set on HO3D [116]. The instances with a small disagreement measure tend to have high validation scores. In contrast, the instances with a high disagreement measure entail false predictions, *e.g.*, detecting the hand-held object as a hand joint and hand class. When the disagreement measure was high at the bottom of Fig. 6.3, we found that both predictions were particularly unstable on the keypoints of the ring finger (yellow). This study shows that the disagreement measure can represent the correctness of the target predictions.

With the disagreement measure ℓ_{disagree}, we define a confidence weight $w_{\mathrm{t}} \in [0, 1]$ for assigning importance to the consistency training. We compute the weight w_{t} as $w_{\mathrm{t}} = 2\left(1 - \mathrm{sigm}\left(\lambda_{\mathrm{d}}\ell_{\mathrm{disagree}}\left(\boldsymbol{\theta}^{\mathrm{tch1}}, \boldsymbol{\theta}^{\mathrm{tch2}}, \boldsymbol{x}_{\mathrm{t}}\right)\right)\right)$ where w_{t} is a normalized disagreement measure with sign inversion, $\mathrm{sigm}(\cdot)$ denotes a sigmoid function, and λ_{d} controls the scale of the measure. With the confidence weight w_{t}, we enforce the consistency training between the student's prediction on the augmented target images $\boldsymbol{p}^k_{\mathrm{s,aug}}$ and the ensemble of the two teachers' predictions $\boldsymbol{p}^k_{\mathrm{ens}}$. Our proposed loss function of confidence-aware geometric augmentation consistency

(C-GAC) $\mathcal{L}_{\text{cgac}}$ for the student $\boldsymbol{\theta}^{\text{stu}}$ is formulated as

$$\mathcal{L}_{\text{cgac}}\left(\boldsymbol{\theta}^{\text{stu}}, \boldsymbol{\theta}^{\text{tch1}}, \boldsymbol{\theta}^{\text{tch2}}, X_{\text{t}}, \mathcal{T}\right) = \mathbb{E}_{\boldsymbol{x}_{\text{t}},(T_{\text{x}},T_{\text{y}}^{\text{p}},T_{\text{y}}^{\text{m}})}\left[w_{\text{t}} \sum_{k\in\{\text{p,m}\}} \tilde{\lambda}^k \tilde{\mathcal{L}}^k \left(\boldsymbol{p}_{\text{s,aug}}^k, T_{\text{y}}^k(\boldsymbol{p}_{\text{ens}}^k)\right) \right], \tag{6.4}$$

where $\boldsymbol{p}_{\text{s,aug}}^k = f^k\left(T_{\text{x}}(\boldsymbol{x}_{\text{t}}); \boldsymbol{\theta}^{\text{stu}}\right)$. Following [268, 329], we design the student prediction $\boldsymbol{p}_{\text{s,aug}}^k$ to be supervised by the teachers. We generate the teachers' prediction by doing ensemble $\boldsymbol{p}_{\text{ens}}^k$, which is better than the prediction of either teacher.

6.3.3 Teacher-student update by knowledge distillation

In addition to the student's training, we formulate an update rule for the two teacher networks by using knowledge distillation. Since ℓ_{disagree} would not work if the two teachers had the same output values, we aim to learn two teachers that have different representations yet keep high task performance as co-training works [30, 56, 275, 296]. In a prior teacher-student update, Tarvainen *et al.* [329] found that the teacher's update by an exponential moving average (EMA), which averages the student's weights iteratively, makes the teacher's learning more slowly and mitigates the confirmation bias as discussed in Sec. 6.3.2. While this EMA-based teacher-student framework is widely used in various domain adaptation tasks [39, 95, 105, 185, 374], naively applying the EMA rule to the two teachers would produce exactly the same weights for both networks.

To prevent this, we propose independent knowledge distillation for building two different teachers. The distillation matches the teacher-student predictions in the output level. To let both networks have different parameters, we train the teachers from different mini-batches and using stochastic augmentation as

$$\mathcal{L}_{\text{distill}}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{stu}}, X_{\text{t}}, \mathcal{T}\right) = \mathbb{E}_{\boldsymbol{x}_{\text{t}},T_{\text{x}}}\left[\sum_{k\in\{\text{p,m}\}} \tilde{\lambda}^k \tilde{\mathcal{L}}^k(\boldsymbol{p}_{\text{t,aug}}^k, \boldsymbol{p}_{\text{s,aug}}^k) \right], \tag{6.5}$$

where $\boldsymbol{\theta} \in \{\boldsymbol{\theta}^{\text{tch1}}, \boldsymbol{\theta}^{\text{tch2}}\}$, $\boldsymbol{p}_{\text{t,aug}}^k = f^k\left(T_{\text{x}}(\boldsymbol{x}_{\text{t}}); \boldsymbol{\theta}\right)$, and $\boldsymbol{p}_{\text{s,aug}}^k = f^k(T_{\text{x}}\left(\boldsymbol{x}_{\text{t}}\right); \boldsymbol{\theta}^{\text{stu}})$. The distillation loss $\mathcal{L}_{\text{distill}}$ is used for updating the teacher networks only. This helps the teachers to adapt to the target domain more carefully than the student and avoid falling into exactly the same predictions on a target instance.

6.3.4 Overall objectives

Overall, the objective of the student's training consists of the supervised loss (Eq. (6.1)) from the source domain and the self-training with confidence-aware geometric augmentation consistency (Eq. (6.4)) in the target domain as

$$\min_{\boldsymbol{\theta}^{\mathrm{stu}}} \mathcal{L}_{\mathrm{task}}\left(\boldsymbol{\theta}^{\mathrm{stu}}, X_{\mathrm{s}}, Y_{\mathrm{s}}\right) + \mathcal{L}_{\mathrm{cgac}}\left(\boldsymbol{\theta}^{\mathrm{stu}}, \boldsymbol{\theta}^{\mathrm{tch1}}, \boldsymbol{\theta}^{\mathrm{tch2}}, X_t, \mathcal{T}\right). \tag{6.6}$$

The two teachers are asynchronously trained with the distillation loss (Eq. (6.5)) in the target domain, which is formulated as

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{distill}}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{stu}}, X_{\mathrm{t}}, \mathcal{T}\right), \tag{6.7}$$

where $\boldsymbol{\theta} \in \{\boldsymbol{\theta}^{\mathrm{tch1}}, \boldsymbol{\theta}^{\mathrm{tch2}}\}$. Since the teachers are updated carefully and can perform better than the student, we use the ensemble of the two teachers' predictions for a final output in inference.

6.4 Experiments

In this section, we first present our experimental datasets and implementation details and then provide quantitative and qualitative results along with the ablation studies. We analyze our proposed method by comparing it with several existing methods in three different domain adaptation settings. We also show qualitative results by applying our method to in-the-wild egocentric videos.

6.4.1 Experiment setup

Datasets: We experimented with several hand datasets including a variety of hand-object interactions, the annotation of 2D hand keypoints, and hand masks as follows. We adopted **DexYCB [49]** dataset as our source dataset since it contains a large amount of training images, their corresponding labels, and natural hand-object interactions. We chose to use the following datasets as our target datasets: **HO3D [116]** captured in different environments with the same YCB objects [42] as the source dataset, **HanCo [419]** captured in a multi-camera studio and generated with synthesized backgrounds, and **FPHA [103]** captured by a first-person view. We also used **Ego4D [111]** to verify the effectiveness of our

method in real-world scenarios. During training, we used cropped images of the hand regions from the original images as input.

Implementation details: Our teacher-student networks share an identical network architecture, which consists of a unified feature extractor and task-specific branches for hand keypoint regression and hand segmentation. For training our student network, we used the Adam optimizer [163] with a learning rate of 10^{-5}, while the learning rate of the teacher networks was set to 5×10^{-6}. We set the hyperparameters ($\lambda^{\mathrm{p}}(= \tilde{\lambda}^{\mathrm{p}})$, λ^{m}, $\tilde{\lambda}^{\mathrm{m}}$, λ_{d}) to $(10^7, 10^2, 5, 0.5)$. Since both task-specific branches have different training speeds, we began our adaptation with the backbone and keypoint regression branch. We then trained all sub-networks, including the hand segmentation branch. We report the percentage of correct keypoints (PCK) and the mean joint position error (MPE) for hand keypoint regression, and the intersection over union (IoU) for hand segmentation.

Baseline methods: We compared quantitative performance with the following methods. **Source only** denotes the network trained on the source dataset without any adaptation. To compare with another adaptation approach with adversarial training, we trained **DANN [102]** that aligns marginal feature distributions between domains, and **RegDA [151]** with an adversarial regressor that optimizes domain disparity. In addition, we implemented several self-training adaptation methods by replacing pseudo-labeling with consistency training. **GAC** is a simple baseline with the consistency training updated by Eq. (6.2). **GAC + UMA [37]** is a GAC method with confidence estimation by Dropout [101]. **GAC + CPL [256]** is a GAC method with confident instance selection using the agreement with another network. **GAC + MT [329]** is a GAC method with the single-teacher-single-student architecture using EMA for the teacher update. **Target only** indicates the network trained on the target dataset with labels, which shows an empirical performance upper bound.

Our method: We denote our full method as **C-GAC** introduced in Sec. 6.3.4. As an ablation study, we present a variant of the proposed method as **GAC-Distill** with a teacher-student pair, which is updated by the consistency training (Eq. (6.2)) and the distillation loss (Eq. (6.5)). **GAC-Distill** is different from **GAC + MT** only in the way of the teacher update.

Method	2D Pose		Seg	2D Pose + Seg
	PCK ↑ (%)	MPE ↓ (px)	IoU ↑ (%)	Avg. ↑ (%)
Source only	42.8/33.5	15.39/19.32	57.9/49.1	50.3/41.3
DANN [102]	49.0/46.8	12.39/13.39	52.8/54.7	50.9/50.8
RegDA [151]	48.8/48.2	12.50/12.64	55.7/55.3	52.2/51.7
GAC	47.6/47.4	12.47/12.54	58.0/56.9	52.8/52.2
GAC + UMA [37]	47.1/45.3	12.97/13.51	58.0/55.0	52.5/50.2
GAC + CPL [256]	48.1/48.1	12.74/12.61	57.2/55.6	52.7/51.8
GAC + MT [329]	45.5/44.4	13.65/14.05	54.8/52.3	50.2/48.3
GAC-Distill (Ours)	49.9/50.4	11.98/11.51	60.7/60.6	55.3/55.5
C-GAC (Ours-Full)	50.3/51.1	11.89/11.22	60.9/60.3	55.6/55.7
Target only	55.1/58.6	11.00/9.29	68.2/66.1	61.7/62.4

Table 6.1: **DexYCB [49] → HO3D [116]**. We report PCK (%) and MPE (px) for hand keypoint regression and IoU (%) for hand segmentation. Each score format of *val* / *test* indicates the validation and test scores. Red and blue letters indicate the best and second best values.

6.4.2 Quantitative results

We show the results of three adaptation settings: DexYCB → {HO3D, HanCo, FPHA} in Tabs. 6.1 and 6.2. We then provide detailed comparisons of our method.
DexYCB → HO3D: Tab. 6.1 shows the results of the adaptation from DexYCB to HO3D where the grasping objects are overlapped. The baseline of the consistency training (**GAC**) was effective in learning target images in both tasks. Our proposed method (**C-GAC**) improved by 5.3/14.4 in the average task score from the source-only performance. The method also outperformed all comparison methods and achieved close performance to the upper bound.
DexYCB → HanCo: Tab. 6.2 shows the results of the adaptation from DexYCB to HanCo across laboratory setups. The source-only network less generalized to the target domain because the HanCo has diverse backgrounds, while **GAC**

Method	DexYCB → HanCo				DexYCB → FPHA			
	2D Pose		Seg		2D Pose		Seg	
	PCK ↑ (%)	MPE ↓ (px)	IoU ↑ (%)	Avg. ↑ (%)	PCK	MPE	IoU	Avg.
Source only	26.0/27.3	21.82/21.48	41.8/41.4	33.9/34.3	14.0	31.32	24.8	19.4
DANN [102]	32.3/33.0	19.99/19.82	56.3/56.9	44.3/45.0	24.4	25.79	28.4	26.4
RegDA [151]	33.0/33.6	19.51/19.44	57.8/58.4	45.4/46.0	23.7	24.27	41.7	32.7
GAC	36.6/37.1	16.63/16.59	58.1/58.8	47.4/47.9	37.2	17.02	33.3	35.3
GAC + UMA [37]	35.1/35.6	17.51/17.48	57.1/57.7	46.1/46.6	36.8	17.29	39.2	38.0
GAC + CPL [256]	32.7/33.5	19.85/19.62	55.8/56.4	44.2/45.0	25.7	24.99	32.7	29.2
GAC + MT [329]	33.2/33.8	18.93/18.83	54.3/55.1	43.8/44.4	31.3	20.81	38.4	34.9
GAC-Distill (Ours)	38.8/39.5	16.06/15.97	57.5/57.7	48.1/48.6	36.8	15.99	35.5	36.1
C-GAC (Ours-Full)	39.2/39.9	15.83/15.74	58.2/58.6	48.7/49.2	37.2	15.36	37.7	37.4
Target only	76.8/77.3	4.91/4.80	75.9/76.1	76.3/76.7	63.3	8.11	-	-

Table 6.2: **DexYCB [49] → {HanCo [419], FPHA [103]}**. We report PCK (%) and MPE (px) for hand keypoint regression and IoU (%) for hand segmentation. We show the validation and test results on HanCo and the validation results on FPHA. Red and blue letters indicate the best and second best values.

succeeded in adapting up to 47.4/47.9 in the average score. Our method **C-GAC** showed further improved results in hand keypoint regression.

DexYCB → FPHA: Tab. 6.2 also shows the results of the adaptation from DexYCB to FPHA, which captures egocentric users' activities. Since hand markers and in-the-wild target environments cause large appearance gaps, the source-only performance performed the most poorly among the three adaption settings. In this challenging setting, **RegDA** and **GAC + UMA** performed well for hand segmentation, while their performance on hand keypoint regression was inferior to the **GAC** baseline. Our method **C-GAC** further improved than the **GAC** method in the MPE and IoU metrics and exhibited stability in adaptation training among the comparison methods.

Comparison to different confidence estimation methods: We compare the results with existing confidence estimation methods. **GAC + UMA** and **GAC + CPL** estimate the confidence of target predictions by computing the variance of multiple stochastic forwards and the task scores between a training network and an

auxiliary network, respectively. **GAC + UMA** performed effectively on DexYCB $\rightarrow$ FPHA, whereas the performance gain was thin in the other settings compared to **GAC**. **GAC + CPL** worked well for keypoint regression on DexYCB $\rightarrow$ HO3D, but it cannot address the other settings with a large domain gap well since the prediction of the auxiliary network became unstable. Although these prior methods had different disadvantages depending on the settings, our method **C-GAC** using the divergence of the two teachers for confidence estimation performed stably in the three settings.

Comparison to standard teacher-student update: We compare our teacher update with the update with an exponential moving average (EMA) [329]. The EMA-based update (**GAC-MT**) degraded the performance from the source only in hand segmentation in Tab. 6.1. This suggests that the EMA update can be sensitive to the task. In contrast, our method **GAC-Distill** matching the teacher-student predictions in the output level did not produce such performance degeneration and worked more stably.

Comparison to adversarial adaptation methods: We compared our method with another major adaptation approach with adversarial training. In Tabs. 6.1 and 6.2, the performance of **DANN** and **RegDA** was mostly worse than the consistency-based baseline **GAC**. We found that instead of matching features between both domains [102, 151], directly learning target images by the consistency training was critical in the adaptation of our tasks.

Comparison to an off-the-shelf hand pose estimator: We tested the generalization ability of an open-source library for pose estimation: **OpenPose** [135]. It resulted in $15.75/12.72$, $18.31/18.42$, and 29.02 in the MPE on HO3D, HanCo, and FPHA, respectively. Since it is built on multiple source datasets [7, 155, 216], the baseline showed higher generalization than the source-only network. However, the performance did not exceed our proposed method in the MPE. This shows that generalizing hand keypoint regression to other datasets is still challenging, and our adaptation framework supports improving target performance.

6.4.3 Qualitative results

We show the qualitative results of hand keypoint regression and hand segmentation in Fig. 6.4. When hands are occluded in HO3D and FPHA or the backgrounds

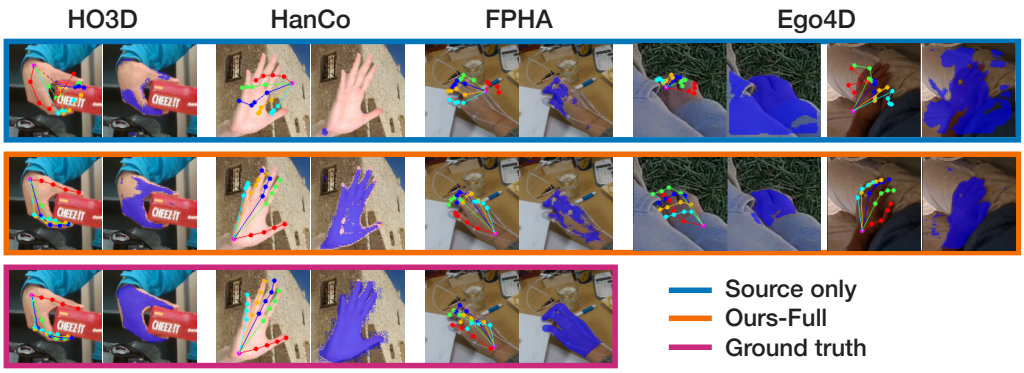


Figure 6.4: **Qualitative results.** We show qualitative examples of the source-only network (top), the Ours-Full method (middle), and ground truth (bottom) on HO3D [116], HanCo [419], FPHA [103], and Ego4D [111] without ground truth.

are diverse in HanCo, the keypoint prediction of the source only (top) represented infeasible hand poses and hand segmentation was too noisy or missing. However, our method **C-GAC** (middle) corrected the hand keypoint errors and improved to localize hand regions. Hand segmentation in FPHA was still noisy because visible white markers obstructed hand appearance. We can also see distinct improvements in the Ego4D dataset.

6.4.4 Ablation studies

Effect of confidence estimation: To confirm the effect of our proposed confidence estimation, we compare our full method **C-GAC** and our ablation model **GAC-Distill** without the confidence weighting. In Tabs. 6.1 and 6.2, while **GAC-Distill** mostly surpassed the comparison methods in most cases, **C-GAC** showed further performance gain in all three adaptation settings.

Multi-task *vs*. single-task adaptation: We studied the effect of our multi-task adaptation compared with single-task adaptation on DexYCB $\rightarrow$ HO3D. The single-task adaptation results are $50.1/51.0$ in the PCK and $58.2/57.7$ in the IoU. Compared to Tab. 6.1, our method in the multi-task setting improved by $2.7/2.6$ over the single-task adaption in hand segmentation while it provided marginal gain in hand keypoint regression. This shows that the adaptation of hand keypoint regression helps to localize hand regions in the target domain.

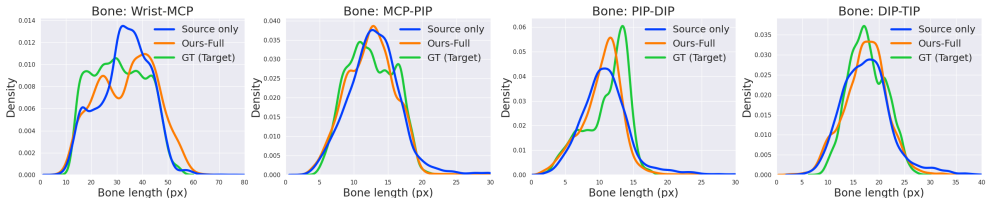


Figure 6.5: **Visualization of bone length distributions.** We show the distributions of the bone length between hand joints, namely, Wrist, metacarpophalangeal (MCP), proximal interphalangeal (PIP), distal interphalangeal (DIP), and fingertip (TIP). Using kernel density estimation, we plotted the density of the bone length for the predictions of the source only, the Ours-Full method, and ground truth on test data of HO3D [116].

Bone length distributions: To study our adaptation results in each hand joint, we show the distributions of bone length between hand joints in Fig. 6.5. In Wrist-MCP, PIP-DIP, and DIP-TIP, the distribution of the source-only prediction on target images (blue) was far from that of the target ground truth (green), whereas our method (orange) improved to approximate the target distribution (green). In MCP-PIP, we could not observe such clear differences because the source-only model already represented the target distribution well. This indicates that our method improved to learn hand structure near the palm and fingertips.

6.5 Additional Results and Details

6.5.1 Dataset details

We provide the details of the datasets used in our experiments.

- **DexYCB [49]** contains 582K RGB-D frames captured by 10 subjects interacting 20 different YCB objects [42] from eight different views. In our experiment, we split the dataset by the subject IDs to create train, validation, and test sets with 212K, 71K, and 80K images, respectively.
- **HO3D [116]** contains 103K RGB-D frames captured by 10 subjects interacting 10 different YCB objects [42] from a single third-person view. In our experiment, we randomly split the video sequences to train, validation, and test sets with 51K, 12K, and 8K images, respectively.
- **HanCo [419]** is an extended FreiHAND [421] dataset captured in a multi-view camera setup with eight cameras, which consists of 518K, 106K, and 104K RGB images for training, validation, and testing, respectively. The backgrounds are randomly synthesized using diverse scenery images.
- **FPHA [103]** is an egocentric video dataset capturing users' actions in daily indoor environments from a first-person perspective, and their hand poses are tracked by hand magnetic sensors. It contains 69K training images and 16K validation images. Due to lacking hand mask annotation, we annotated 50 hand masks in the validation set.
- **Ego4D [111]** is a collection of daily-life egocentric activity videos lasting over 3,000 hours and gathered across the world. Due to the lack of annotation for the two tasks, we show qualitative examples in our experiments. We treated each video sequence as the domain to adapt.

6.5.2 Preprocessing and augmentation

For creating an input of a training network, we assumed to have hand center positions, cropped hand regions of the original images, and resized them to 128×128 pixels. To extract hand centers and regions in Ego4D videos without ground truth,

we used an off-the-shelf hand detector [305]. Inspired by [79, 185, 202], we used two different augmentation sets: strong augmentation for the student's learning (Eq. (6.4)) and weak augmentation for the teacher's learning (Eq. (6.5)). We used horizontal flip, rotation, transition, gaussian blur, brightness/contrast jitter, hue/saturation/input value jitter, and cutout as the strong augmentation. In contrast, we adopted horizontal flip, rotation, transition, and gaussian blur as the weak augmentation.

6.5.3 Network architecture and evaluation

For the design of our multi-task baseline model, we employed an hourglass network [242] as the backbone and the keypoint regression branch. We added 1d-convolution to its intermediate features to predict hand pixel labels. Following hand keypoint regression methods [104, 242, 358], we optimized 2D joint heatmaps for each 2D ground-truth joint location instead of joint coordinates.

We also provide the details of out evaluation, namely, MPE, PCK, and IoU. MPE (px) indicates the euclidean error per joint in the image coordinate. PCK (%) represents the percentage of joints whose MPE is smaller than a given joint error threshold, which is calculated by the area under the curve (AUC) over the joint error range [0, 20 px]. IoU (%) measures the overlap over two masks. We report the average score (Avg.) over PCK and IoU to evaluate multi-task performance.

6.5.4 Additional qualitative results

In Figs. 6.6, 6.7, 6.8, and 6.9, we show additional qualitative results of our proposed method. As shown in Fig. 6.6, our method performed well when complex hand-object interactions occur on HO3D and FPHA and when the backgrounds are diverse on HanCo. In Fig. 6.7, we show qualitative comparison between GAC and C-GAC (Ours-Full). Our full method particularly improved keypoint regression compared to the simple consistency baseline, GAC. Our method (right) corrected the keypoint prediction of the GAC (left), which contains incorrect predictions on the position of the thumb (red).

Our method also demonstrated improved performance on Ego4D, an egocentric video dataset collected across various countries, cultures, ages, indoors/outdoors,

and performing tasks with hands. In particular, we observed that our method successfully adapted to various imaging conditions, such as outdoor environments (rows 1 and 2 in Fig. 6.8), extremely dark environments (rows 3 to 6 in Fig. 6.8), the second person's hands in social interactions (row 7 in Fig. 6.8), *e.g.*, playing board games, and indoor environments (Fig. 6.9), *e.g.*, where people perform cooking, cleaning, fitness, DIY, painting, and crafting.

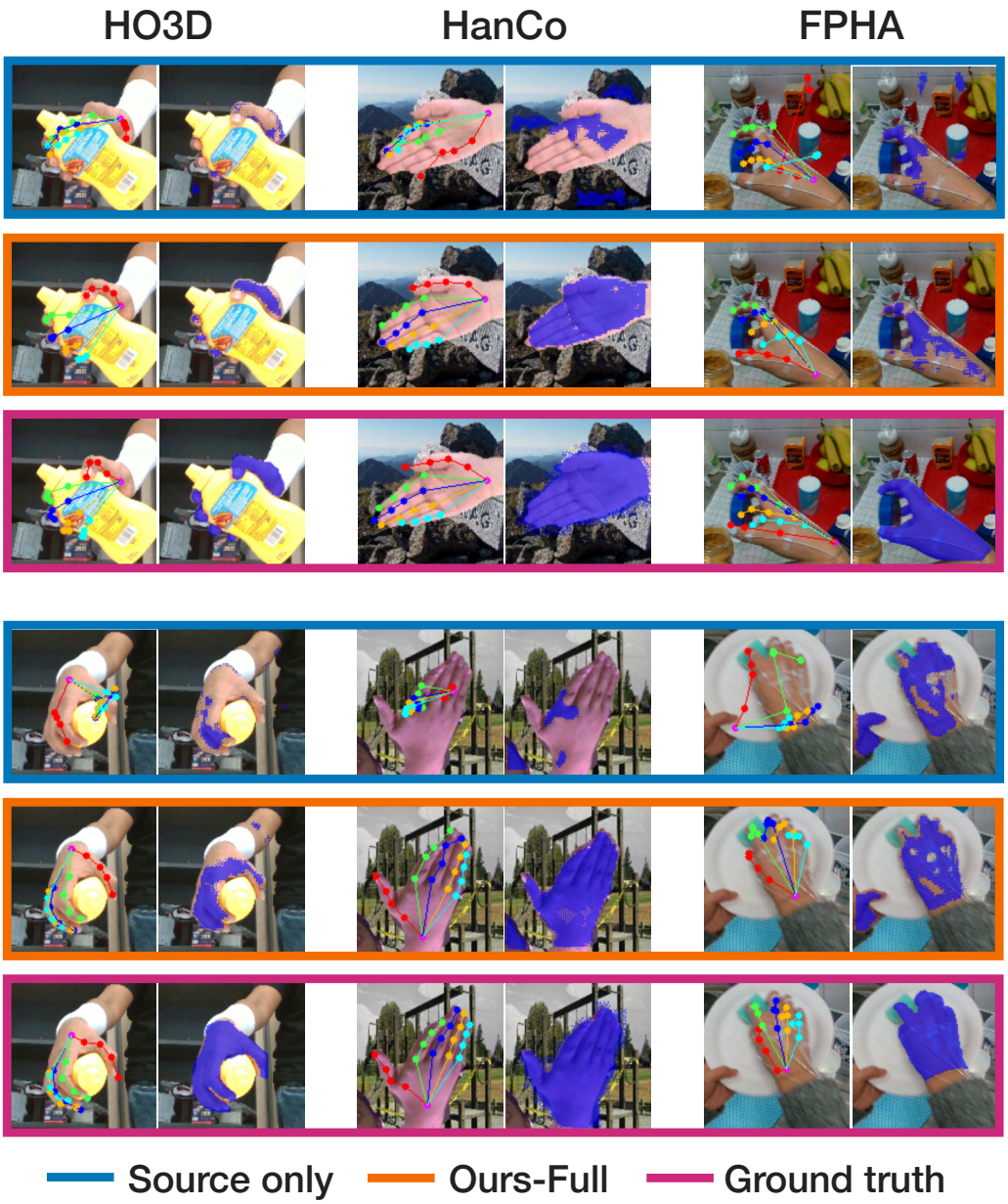


Figure 6.6: **Additional qualitative results on HO3D [116], HanCo [419], and FPHA [103].**

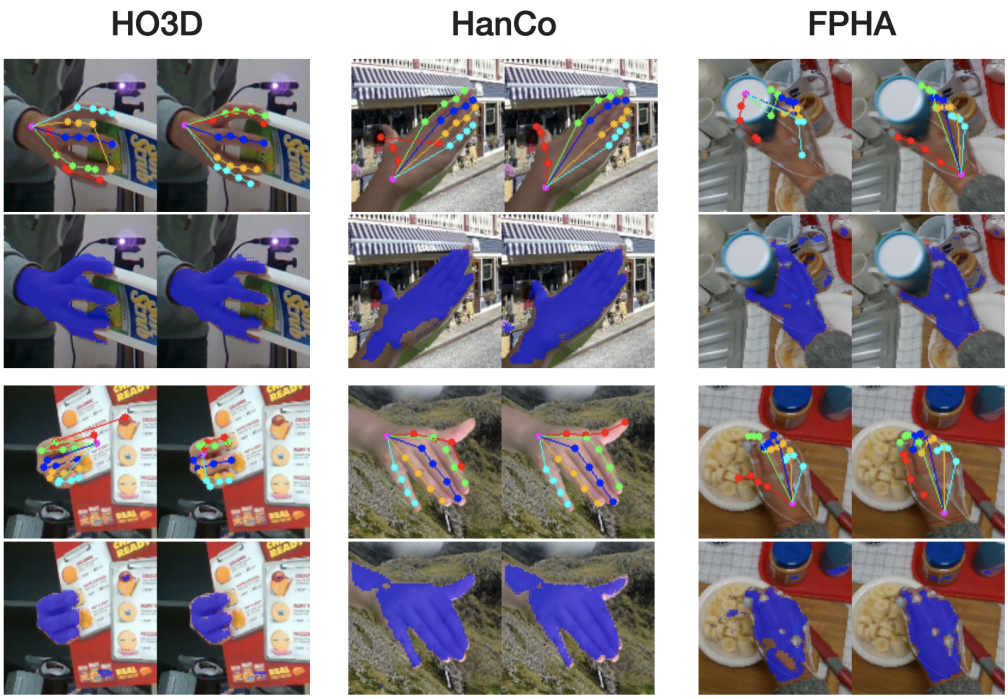


Figure 6.7: **Comparison between GAC and C-GAC (Ours-Full).** Left: GAC, Right: C-GAC (Ours-Full).

Figure 6.8: **Additional qualitative results on Ego4D [111].**

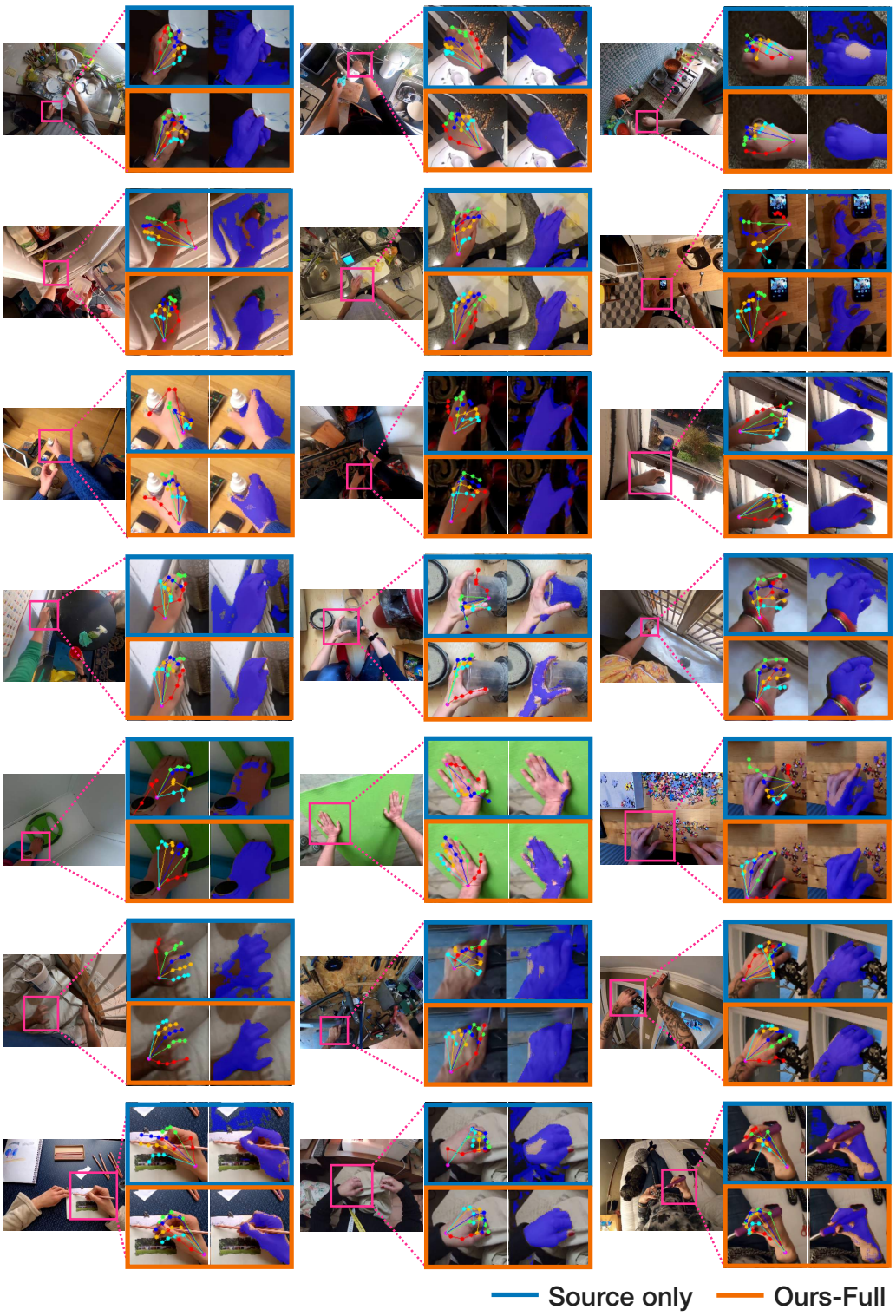


Figure 6.9: **Additional qualitative results on Ego4D [111].**

6.6 Conclusion

In this work, we tackled the problem of joint domain adaptation of hand keypoint regression and hand segmentation. Our proposed method consists of the self-training with geometric augmentation consistency, confidence weighting by the two teacher networks, and the teacher-student update by knowledge distillation. The consistency training under geometric augmentation served to learn the unlabeled target images for both tasks. The divergence of the predictions from two teacher networks could represent the confidence of each target instance, which enables the student network to learn from reliable target predictions. The distillation-based teacher-student update guided the teachers to learn from the student carefully and mitigated over-fitting to the noisy predictions. Our method delivered state-of-the-art performance on the three adaptation setups. It also showed improved qualitative results in the real-world egocentric videos.

Chapter 7

Dense Video Captioning for Egocentric Hand Activities

In this chapter, we propose a novel benchmark for cross-view knowledge transfer of dense video captioning, adapting models from web instructional videos with exocentric views to an egocentric view. While dense video captioning (predicting time segments and their captions) is primarily studied with exocentric videos (*e.g.*, YouCook2), benchmarks with egocentric videos are restricted due to data scarcity. To overcome the limited video availability, transferring knowledge from abundant exocentric web videos is demanded as a practical approach. However, learning the correspondence between exocentric and egocentric views is difficult due to their dynamic view changes. The web videos contain shots showing either full-body or hand regions, while the egocentric view is constantly shifting. This necessitates the in-depth study of cross-view transfer under complex view changes. To this end, we first create a real-life egocentric dataset (EgoYC2) whose captions follow the definition of YouCook2 captions, enabling transfer learning between these datasets with access to their ground-truth. To bridge the view gaps, we propose a view-invariant learning method using adversarial training, which consists of pre-training and fine-tuning stages. Our experiments confirm the effectiveness of overcoming the view change problem and knowledge transfer to egocentric views. Our benchmark pushes the study of cross-view transfer into a new task domain of dense video captioning and envisions methodologies that describe egocentric videos in natural language.

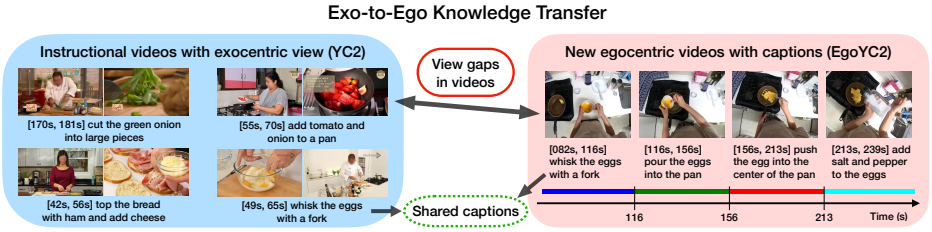


Figure 7.1: **Our cross-view knowledge transfer of dense video captioning.** We propose to utilize existing web instructional videos with exocentric views, YouCook2 (YC2) [407], to improve dense video captioning on newly recorded egocentric videos (EgoYC2). The EgoYC2's captions are annotated by following YC2, enabling the study of transfer learning under view gaps in videos.

7.1 Introduction

Perceiving procedural human activities from an egocentric (first-person) view has been a long-standing problem. Compared to action recognition focusing on labeling specific activities [77, 303], video-to-text description extends this realm, offering a detailed textual interpretation of ongoing activities. This not only facilitates understanding of the task procedure, but enriches intuitive and communicative interfaces between humans and assistive machine systems, *e.g.*, augmented reality [212] and human-robot interactions [156].

One formulation of video-to-text description is dense video captioning [168], which densely detects time segments of a video and generates their captions. This task has been studied with instructional videos (*e.g.*, YouCook2 [407]). While these instructional videos primarily feature exocentric (third-person) views available on the Web, egocentric benchmarks for dense video captioning remain underexplored due to limited dataset scale (*e.g.*, 16 hours of MMAC Captions (ego) [239] *vs*. 176 hours of YouCook2).

Given this data scarcity, finding ways to bridge exocentric (source) and egocentric (target) views is vital to utilizing numerous web exocentric videos to enhance the understanding of egocentric activities. Prior works of this cross-view transfer have been studied in action recognition [186, 352] and human/hand pose estimation [159, 252].

Unlike short-term modeling in these prior works (*e.g.*, estimation of per-frame

poses and per-clip actions), dense video captioning requires addressing longer sequence inputs to model the coherence of each step description. This emphasizes the problem of dynamic camera view changes in learning the correspondence between both views. The source instructional videos are not purely captured from a single fixed camera, but composed of multiple alternating views (*e.g.*, face and hand shots [224]). In contrast, egocentric videos inherently include dynamic view changes due to the camera wearer's motion, which obstructs learning the procedure [20]. This necessitates adapting captioning models from mixed source views to a moving target view.

In this work, we propose a knowledge transfer benchmark for dense video captioning from web instructional videos with exocentric views to egocentric videos, with access to their ground-truth during training (Fig. 7.1). To study this cross-view transfer, we first create a new egocentric dataset, EgoYC2, with caption annotations following a source dataset, YouCook2. We collect 226 egocentric cooking videos from 44 users, featuring real-life home kitchens (*vs*. a laboratory kitchen in MMAC Captions). These paired datasets reduce discrepancies in caption content and granularity, allowing us to pre-train a model on the source data and fine-tune it on the target data.

To address the view gaps, we encourage pre-training and fine-tuning to be less affected by view-dependent bias, using adversarial invariant feature learning [57, 102]. The pre-training aims to learn features invariant to the two views in the source data: face shots showing body actions and hand shots focusing on hand-object interactions. The view-invariant fine-tuning is performed using the source and target datasets, further adapting to the egocentric domain. Additionally, we observe camera motion in egocentric videos intensifies the view gaps; thus we stabilize the videos by a fine and temporally coherent tracking of hand-object interactions, which consists of hand detection and tracking [26, 305] and hand-object segmentation [400].

We evaluate our transfer learning method regarding how effectively the method overcomes view changes and efficiently transfers knowledge to the egocentric domain. Our pre-training with the decoupling of mixed source views has shown further improvement in egocentric video captioning against a naive pre-trained model. The stabilization of the target view movement with hand tracking improves fine-tuning performance, and additional support of hand-object features is

more effective. Our view-invariant fine-tuning further improves adapting the pre-trained model to the egocentric domain. Our benchmark allows us to provide a practical solution for transfer learning from exocentric to egocentric videos under dynamic view changes.

Our contributions are summarized as follows:

- We offer a new real-life egocentric video dataset (EgoYC2) for dense video captioning, whose captions follow those of exocentric videos (YouCook2).
- We propose view-invariant learning in pre-training and fine-tuning with unified adversarial training and video processing that mitigates the view change effects.
- We demonstrate how effectively the proposed method overcomes the problem of view changes and efficiently transfers the knowledge to the egocentric domain.

7.2 Related Work

Dense video captioning: Dense video captioning consists of two sub-tasks localizing multiple time segments occurring in a video and describing their captions. While Krishna *et al*. initially proposes to describe coarse activities [168], subsequent works focus on more fine-grained activities using instructional videos, such as cooking [239, 407] (*e.g*., YouCook2 (YC2)), makeup [356], and daily activity videos [4, 222, 356]. These instructional videos promote the understanding of the *procedure* [407], a series of steps to accomplish certain tasks.

However, dense video captioning for egocentric (first-person) videos has been less studied compared to exocentric (third-person) benchmarks [4, 222, 327, 356, 407]. A video captioning dataset for egocentric cooking videos, MMAC Captions [239], contains x11 smaller amount of videos than YC2. Other related egocentric datasets (*e.g*., EPIC-KITCHENS [77] and Ego4D [111]) are collected with spoken narrations. Their transcribed texts help define action labels and video-language pre-training [193, 272], but they drastically differ from *procedural* captions [375, 376]. In addition, its supervision is typically weak and noisy, including incorrect visual grounding, irrelevant caption content, and errors in automatic

speech recognition [166, 195, 221, 315]. Xu *et al*. [369] propose to use LLM to refine the noisy captions, enabling egocentric captioning with retrieval.

We newly construct an egocentric video dataset (EgoYC2) with *procedural* caption annotations. Compared to the MMAC Captions dataset captured in a laboratory kitchen, we record real-life cooking activities and annotate captions following the exocentric dataset YC2. This allows us to let models resolve the view gaps only without considering the gaps in caption content and granularity. In fact, we use the same vocabulary list as the YC2 and have close average step sizes per video (7.7 in YC2 *vs*. 6.5 in EgoYC2 *vs*. 30.1 in MMAC Captions).

Egocentric perception using exocentric videos: Exocentric view data can inform the state of humans, actions, and the surrounding environment that are not always observable from a limited field-of-view of an egocentric camera. Thus, knowledge transfer from exocentric to egocentric views is essential to complement egocentric perception, *e.g*., in action recognition [186, 308, 334, 352, 372] and human/hand pose estimation [159, 252, 255]. This transfer has been categorized into two settings: *paired vs*. *unpaired* scenarios.

The *paired* setting assumes that the same actions are captured from different views [308] or using synchronized egocentric and exocentric cameras [112, 171, 303]. Sigurdsson *et al*. propose to learn a shared feature space between both views using metric learning [308]. With camera calibration and head-camera tracking, other works [159, 171, 252] project annotated hand-object poses from exocentric to egocentric views. The *unpaired* setting relaxes this assumption of view correspondence; thus, two videos are neither synchronized nor captured in the same environment. Several ways to learn view-invariant features in unpaired videos have been proposed, such as knowledge distillation [186] and cross-view feature alignment [352, 372] or attention [334]. Other methods employ domain adaptation techniques, such as adversarial training [70, 238] and pseudo-labeling [255, 256].

Our work addresses this cross-view transfer problem for a new task domain, dense video captioning. Unlike action recognition, the transfer of dense video captioning requires handling longer sequence inputs, which highlights view changes as a longer sequence includes various views. Since our work is based on asynchronous videos, we propose an adversarial training method that works in the *unpaired* setting. Compared to adversarial adaptation in action recognition directly bridging between two datasets [57, 70, 238], we follow the idea of gradual

domain adaptation [170, 196, 349] by splitting a large domain gap caused by different recording setups and distinct viewpoints into smaller gaps, and resolve the gaps step-by-step. We offer a practical application of gradual adaptation beyond conventional setups [170, 196, 349], *e.g.*, digit and portrait data.

7.3 Exo-to-Ego Transfer Learning

We propose a solution for a transfer learning task of dense video captioning from web instructional videos to egocentric videos, *i.e.*, YouCook2 (source) → EgoYC2 (target). Given the problem of the limited data scale for egocentric video captioning, our proposed method is motivated to utilize external video resources collected on the Web. We assume that web instructional videos with (dense) captions covering activity classes of egocentric videos are available, and the target ground-truth is accessible, which follows supervised domain adaptation [232, 336]. We first introduce our setup and transfer learning method in Secs. 7.3.1 and 7.3.2, then describe our model with respect to the representation of the hand-object in Sec. 7.3.3 and a captioning network in Sec. 7.3.4.

The challenge of this task is to address the dynamics of view changes in both datasets. Unlike cross-view transfer in action recognition [186, 308, 334, 352, 372], dense video captioning requires modeling longer video sequences, which further complicates the view change problem. We observe that the source web videos can be decomposed into several shots captured from different views, based on video composition analysis, *e.g.*, instruction classes in how-to videos [381], scene cut categorization [224, 263], temporal shot segmentation [58, 361]. In contrast, egocentric videos are typically untrimmed and captured from a single camera, but involve dynamic scene changes due to the head movement. This dynamic movement prevents learning the procedure (key steps) [20]. Thus, given the problems of the mixed and moving views, it is necessary to overcome view-dependent bias by adapting models to the target egocentric domain.

Preliminary: We have access to both labeled source and target datasets $\mathcal{D}_{\mathrm{s}}$ and $\mathcal{D}_{\mathrm{t}}$. These datasets contain labeled videos with the size of m (n), where the input $\mathbf{X}$ is features for a video and $\mathbf{y}$ corresponds to its ground-truth of dense video captions, written as $\mathcal{D}_{\mathrm{s}} = \{(\mathbf{X}_{\mathrm{s}1}, \mathbf{y}_{\mathrm{s}1}), (\mathbf{X}_{\mathrm{s}2}, \mathbf{y}_{\mathrm{s}2}), \ldots, (\mathbf{X}_{\mathrm{s}m}, \mathbf{y}_{\mathrm{s}m})\}$ and $\mathcal{D}_{\mathrm{t}} = \{(\mathbf{X}_{\mathrm{t}1}, \mathbf{y}_{\mathrm{t}1}), (\mathbf{X}_{\mathrm{t}2}, \mathbf{y}_{\mathrm{t}2}), \ldots, (\mathbf{X}_{\mathrm{t}n}, \mathbf{y}_{\mathrm{t}n})\}$. The video features $\mathbf{X}$ are encoded

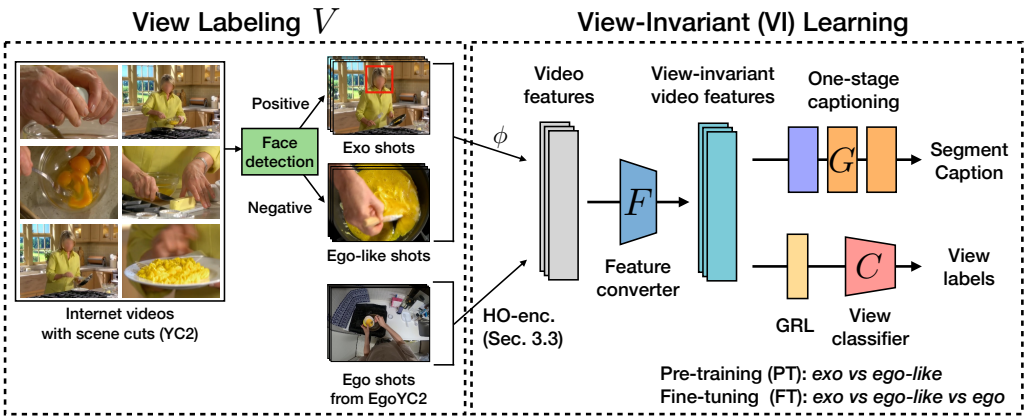


Figure 7.2: **View-invariant learning across exocentric and egocentric views.** (i) We define an intermediate view (*ego-like*) in the source domain, which represents the one between *exo* and *ego* views. We treat source images where the face is detected as the *exo* view and the others as the *ego-like* view due to its similarity to the *ego* view. We generate video features using a fixed encoder ϕ and describe this processing for egocentric videos in Sec. 7.3.3. (ii) We design our view-invariant (VI) learning to gradually adapt from *exo* to *ego* views. Our method consists of pre-training (PT) on the source data and fine-tuning (FT) across the source and target data. Following adversarial domain adaptation [102], we train a feature converter F and a view classifier C with a gradient reversal layer (GRL). This encourages feature learning invariant to the view classes to be classified by C. The former PT takes the source data with the *exo* and *ego-like* classes, while the latter FT takes all views to align them.

by a fixed feature extractor ϕ and represented as a set of frame-wise features: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, where $\mathbf{x}$ is frame features and T is the fixed length of a video.

Unlike a standard formulation of classification tasks [57, 70, 161, 238], dense video captioning poses a complex form of architecture and loss function [354, 355, 408]. Here we encapsulate the captioning model G and task loss $\mathcal{L}_{\text{task}}$, and describe their details in Sec. 7.3.4. Our inference model consists of a feature converter network F and a task network G, learning the mapping from $\mathbf{X}$ to $\mathbf{y}$. We also define the view class for each frame with a view labeling function $V(\cdot)$, which takes a frame $\mathbf{x}$ and assigns it to one of the predefined view labels (*i.e.*,

exo, *ego-like*, and *ego* as indicated in Sec. 7.3.1). Given an entire model $G \circ F$ and a dataset $\mathcal{D}$, we denote a task loss function for dense video captioning as $\mathcal{L}_{\text{task}}(F, G, \mathcal{D})$. Using the source and target datasets, the joint training on the two datasets is defined as

$$\mathcal{L}_{\text{task}}(F, G, \mathcal{D}_{\text{t}}) + \lambda_{\text{src}} \mathcal{L}_{\text{task}}(F, G, \mathcal{D}_{\text{s}}), \tag{7.1}$$

where λ_{src} is a weight controlling the training on $\mathcal{D}_{\text{s}}$.

7.3.1 View labeling and preprocessing

We construct the view labeling function $V(\cdot)$. Based on the composition analysis of exocentric cooking videos [224], the source data can be divided into face shots showing full-body actions and hand shots indicating the necessary attention to specific objects. These shots are interleaved even in several seconds and changing throughout the video sequence. Observing the visual similarity between the hand shots and the egocentric images, we define three view classes as *exo* (face shots) and *ego-like* (hand shots) from the source exocentric videos, and *ego* views from the target egocentric videos (see the left of Fig. 7.2).

For the source data, we use face tracking that classifies the face and hand shots, consisting of a face detector [302] and a tracking method SORT [26]. We assign positive images with face to the *exo* view and the rest to the *ego-like* view. Compared to classifier-based scene categorization in Ego-Exo [186], our classification based on local facial features is more versatile. Li *et al*. [186] employ an ego-exo scene classifier trained on Charades-Ego [308] to assign soft labels between *ego* and *exo* views. This scene classifier may exhibit bias due to its training on mostly side views in [308]'s *exo* views, over the front views prevalent in YC2's *exo* views. Such biases could limit its generalization to unseen videos like EgoYC2 and YC2. In contrast, our face detection-based approach can effectively handle various face angles, regardless of side and front views.

To reduce the moving impact of the *ego* view, we propose using a fine-grained and temporally coherent tracking of hand-object interactions, informing actions occurring in moving scenes. We implement this tracking by using the combination of hand detection-based tracking and frame-wise hand-object segmentation. Specifically, we track the location of bounding boxes covering two hands using a

pre-trained hand detector [305, 328] and the SORT algorithm [26]. Additionally, we use the pixel-level location of hands and interacting objects [400], which can provide more precise descriptions of actions. These two techniques complement each other; the hand tracking is temporally coherent but coarse localization and the segmentation is fine-grained but frame-wise localization (see Sec. 7.3.3 for details).

7.3.2 Transfer learning via view-invariant learning

We aim to learn view-invariant features among the mixed source views and the unique egocentric view. We perform pre-training and fine-tuning separately to handle a larger domain gap. Compared to prior action domain adaptation, our transfer learning will suffer from a larger domain gap as the source and target data are neither constructed within the same dataset [238] nor captured from similar viewpoints [57]. Here we employ a "divide-and-conquer" approach by breaking down a large domain gap into smaller gaps and adapting between the smaller gaps step-by-step, dubbed *gradual* domain adaptation [170, 196, 349]. Specifically, we introduce an intermediate domain in the source data that shares visual similarity with the target data. This encourages us to resolve the large gap *gradually* with separated training stages.

We facilitate the learning of invariant features with adversarial training, while gradually adapting from *exo* to *ego-like* and finally to *ego* view. Fig. 7.2 illustrates the overview of the view labeling and our learning method. The video features are converted to learnable representation by the converter F. These features are fed to the task network G to solve the captioning task and the classifier C to let the converted features be view-independent. The classifier C is trained by adversarial adaptation with a gradient reversal layer [102]. The converter F attempts to produce features undistinguished by the classifier while the classifier is trained to classify their views.

View-invariant pre-training in source domain: We pre-train the model on the source data to let the feature converter F produce view-invariant features. This learning is facilitated by adversarial loss [102] with the classifier C, which learns the mapping from a frame $\mathbf{x}$ to a view label $V(\mathbf{x})$:

$$\mathcal{L}_{\text{adv}}\left(F, C, \mathcal{D}, V\right) = \mathbb{E}_{\mathbf{x}\sim X\sim\mathcal{D}}\mathcal{L}_{\text{ce}}\left(C\left(F\left(\mathbf{x}\right)\right), V\left(\mathbf{x}\right)\right). \tag{7.2}$$

We use a cross-entropy loss $\mathcal{L}_{\text{ce}}$ as the objective to reuse $\mathcal{L}_{\text{adv}}$ in the three-view classification problem in the next section. The total loss of the pre-training is defined as

$$\mathcal{L}_{\text{task}}\left(F, G, \mathcal{D}_{\text{s}}\right) - \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}\left(F, C, \mathcal{D}_{\text{s}}, V\right), \tag{7.3}$$

where λ_{adv} is a controlling weight for the adversarial loss.

View-invariant fine-tuning across three views: We follow similar view-invariant learning during the training with the source and target data. We first initialize the entire model $G \circ F$ from a pre-trained checkpoint (Eq. (7.3)) and reinitialize the classifier C. Then, we jointly train the model on the source and target data (Eq. (7.1)), while the classifier C is trained to classify a frame $\mathbf{x}$ from both datasets into the three view classes as

$$\begin{aligned}\mathcal{L}_{\text{task}}\left(F, G, \mathcal{D}_{\text{t}}\right) + \lambda_{\text{src}}\mathcal{L}_{\text{task}}\left(F, G, \mathcal{D}_{\text{s}}\right) \\ - \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}\left(F, C, \{\,\mathcal{D}_{\text{s}}, \mathcal{D}_{\text{t}}\,\}, V\right).\end{aligned} \tag{7.4}$$

Note that the source web data are typically much larger than recorded egocentric videos ($m \gg n$). To address this, we employ an undersampling technique for task loss of Eq. (7.1), where the number of input source data is balanced to the number of target data per iteration. Similarly, we address the imbalance of view classes in the adversarial loss of Eq. (7.2) with the undersampling.

7.3.3 Hand-object feature generation for model input

We describe feature encoding for egocentric videos. As discussed in Sec. 7.3.1, we crop the videos with hand tracking and utilize hand-object masks to localize actions in moving scenes. Not only reducing the moving effect, these hand and object cues improve recognition of action verbs (*e.g.*, "pour") and objects (*e.g.*, "olive oil").

Estimates of hand-object interactions have been considered as a means of egocentric video representation, *e.g.*, hand masks [187], hand poses [89, 252, 359], and 3D hand-object features [171, 330]. Due to the recent advancement of diverse object segmentation [165], our work utilizes hand tracking and hand-object segmentation for video representation.

From the affordance analysis [78, 388, 400], hand-object interactions are classified into direct and indirect interactions. Direct (first-order) interaction refers to

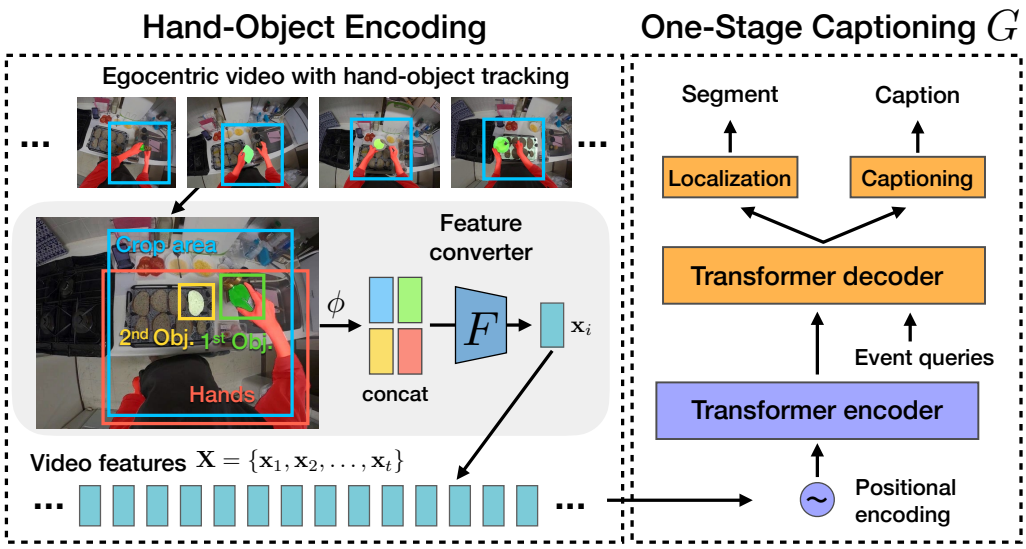


Figure 7.3: **Baseline for egocentric dense video captioning.** Our baseline consists of (i) hand-object encoding and (ii) one-stage captioning with parallel decoding (PDVC [354]). We first preprocess the egocentric videos with hand detection ("crop area") and hand-object segmentation ("hands", "1st obj.", and "2nd obj."). We extract features for these regions by the fixed encoder ϕ and pass their concatenated features to the feature converter F. The generated video features are fed to a transformer-based captioning model with two prediction heads of time segment and caption.

the case where the hand is in physical contact with an object. Indirect (second-order) interaction describes an object while being manipulated using a tool without direct contact. In the scene of "pouring olive oil to bread" in Fig. 7.3, the hand *directly* grasps an olive oil bottle (*i.e.*, first-order object) and pours it onto bread (*i.e.*, second-order object) where the person *indirectly* interacts with the bread.

To recognize these objects, we propose a practical refinement scheme of hand-object masks using two segmentation models. We first adopt EgoHOS [400], a pre-trained segmentation model for hands and interacting objects in egocentric videos. This model tends to produce inaccurate masks for the objects, as detecting object boundaries is difficult in real-life scenes. To correct this error, we use a generic segmentation model for any objects (SAM [165]), which can segment objects in any category and generalize to real-world scenes. Using this capability,

we take an intersection between the two models' predictions and use the most overlapped mask from the SAM for the final output.

Given the hand-object masks, we extract features for regions that enclose each entity with the pre-trained encoder ϕ, namely cropped area (blue), hands (red), first/second-order interacting objects (green/yellow) (see Fig. 7.3). The feature converter F takes their concatenated features and produces frame-wise features for an entire video.

7.3.4 Captioning baseline

We employ a one-stage captioning model to construct the task network G, which has a unified architecture for the two sub-tasks, predicting time segments (*e.g.*, [033s-125s]) and captions (*e.g.*, "apply olive oil and cheesy sauce on the bread"). The primary reason for this choice is that our proposed transfer learning is designed to learn view-invariant features in a middle layer of the encoding pipeline. Thus, using models with separate encoding for the two subtasks would complicate this view-invariant learning. Specifically, we adopt a strong baseline for dense video captioning with parallel decoding (PDVC) [354] (see the right of Fig. 7.3). This model has shown superiority over the two-stage models with the separate encoding scheme [355, 408]. The task loss $\mathcal{L}_{\text{task}}$ of PDVC consists of the localization loss of time segments, the classification for event query, the cross-entropy for predicted words, and the cross-entropy for event count.

Dataset	view	paired?	text?	proc. caption?	domain	source
ActivityNet Captions [168]	exo*		✓		various	YouTube
YouCook2 (YC2) [407]	exo*		✓	✓	cooking	YouTube
EPIC-KITCHENS [77]	ego		✓		cooking	recorded
MMAC Captions [239]	ego		✓	✓	cooking	recorded
YouMakeup [356]	exo*		✓	✓	makeup	YouTube
COIN [327]	exo*		✓	✓	various	YouTube
HowTo100M [222]	exo*		✓		various	YouTube
HIREST [393]	exo*		✓	✓	various	YouTube
Ego4D [111]	ego		✓		various	recorded
Charades-Ego [308]	both	P*			various	recorded
H2O [171]	both	P			various	recorded
Assembly101 [303]	both	P			assembly	recorded
(3+1)ReC [300]	both	P			cooking	recorded
Ego-Exo4D [112]	both	P			various	recorded
EgoYC2 (Ours)	ego	WP	✓	✓	cooking	recorded

Table 7.1: **The comparison of datasets for human activity understanding.** We show the view type ("ego" or "exo") and whether their views are paired ("P": paired, "WP": weakly paired). We compare the presence of textual annotations and whether they take the form of procedural captions [407]. The last two columns indicate the domain and the source of the videos.

Datasets	YC2	EgoYC2
#video	2,000	226
#user	-	44
#recipe_class	89	21
total duration	176 h	43 h
avg. video dur.	5.3 min	11.6 min
avg. segment dur.	19 sec	103 sec
avg. step size	7.7	6.5
viewpoint	exo*	ego
user consent	×	✓

Table 7.2: **Statistics of YouCook2 (YC2) and Ego-YouCook2 (EgoYC2).** We re-record 11.3% of YouCook2 recipes with a head-mounted camera, resulting in 43 hours of 226 videos.

7.4 EgoYC2 Dataset

We describe the details of newly collected EgoYC2. To provide a sound evaluation of knowledge transfer between different datasets, EgoYC2 captions follow the caption definition of YouCook2 (YC2) [407]. This ensures that the two datasets are uniform in caption content and granularity, and are evaluated consistently. Specifically, we directly adopt the *procedural* captions from YC2, describing the sequence of necessary steps to complete complex tasks. We then re-record these cooking videos by instructing participants wearing a head-mounted camera to cook while referring to the YC2's captions (recipes).

The dataset comparison is shown in Tab. 7.1 [1]. Compared to relevant datasets [77, 111, 171, 303, 308], our EgoYC2 dataset has unique features in terms of textual annotations and their quality, and pairing to an external dataset.

[1] "exo*" view indicates that the video may not be captured from a fixed viewpoint, *e.g.*, YouTube videos contain scene cuts with different views.

"P*" in Charades-Ego indicates paired data capturing the same person but not synchronized.

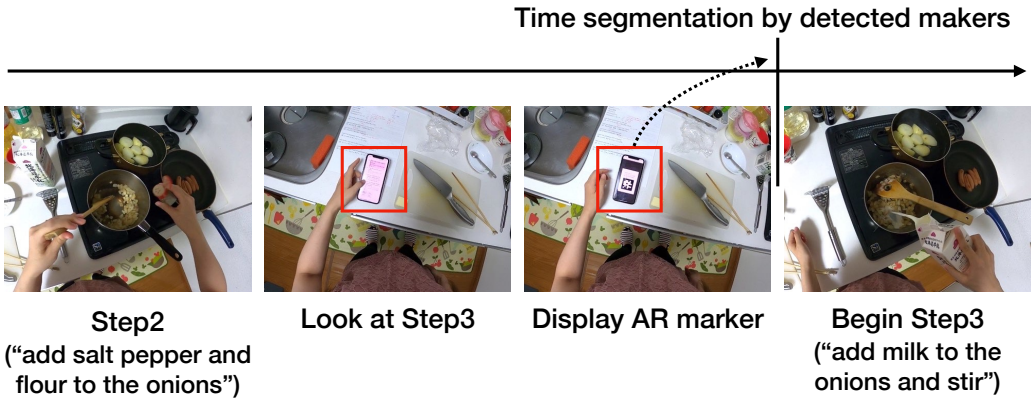


Figure 7.4: **Time segmentation by detected AR markers.** In the transition of cooking steps, we ask participants to check the next step on their smartphone or tablet and display an AR marker once they confirm the next step. Given a recorded video, we postprocess it to detect the marker and segment the video temporally.

Instead of using action labels [171, 303, 308], we newly provide *procedural* captions. Most ego-exo datasets [112, 171, 303] assume camera synchronization between egocentric and exocentric views, while it is laborious to set up. In contrast, our benchmark is based on a further relaxed assumption: *weakly paired* setting where different users could perform actions in different recording setups, but the annotated captions follow the same definition between the two datasets. Furthermore, dense video captioning has not been proposed in popular egocentric video datasets, such as EPIC KITCHENS [77] and Ego4D series [111, 112], and their annotated narrations differ from our *procedural* captions [375, 376].

Regarding knowledge transfer setups, our benchmark is based on a widely applicable assumption: *cross-dataset* transfer, while prior works address the in-dataset transfer, such as Charades-Ego's cross-view training [308] and EPIC-KITCHENS' domain adaptation [238]. This *cross-dataset* assumption does not require strict alignment of the video capture setup between the source and the target data.

Tab. 7.2 shows the statistics compared to YC2. Our videos reach 11.3 % of YC2 videos downloaded from YouTube. Our untrimmed videos have longer video and per-segment duration. Unlike the MMAC Captions [239], we provide caption annotations following the YC2 to let models address the view gaps without con-

sidering caption gaps. We use the same vocabulary list as YC2 and have close step sizes (7.7 in YC2 *vs*. 6.5 in EgoYC2 *vs*. 30.1 in MMAC). We obtain user consent for data collection and release.

Time segment annotation: Besides the annotations of step descriptions, we need to annotate the time segment of each step description (*i.e*., start and end time). We propose automatic time stamp annotation using AR markers displayed on a virtual screen, as shown in Fig. 7.4. We ask participants to use their smartphone or tablet to see a step description in the transition of the steps. We show an AR marker on the screen once they confirm the next step. We postprocess a recorded video by detecting the displayed markers, and segment it temporally. This allows us to annotate the time segments without manually inspecting the entire video.

7.5 Experiments

7.5.1 Experimental setup

Implementation details: We employ PyTorch for implementation and run all experiments on a single NVIDIA V100 16GB GPU. The video features are generated by a pre-trained encoder ϕ, ResNet152 [131]. We present different video presentations for egocentric videos: raw video features (V), cropped video features (VC), and the one with hand-object features (VC+HO). We train models jointly on the source and target data to avoid overfitting the target data. We set λ_{src} and λ_{adv} as 0.1, and the fixed video length T as 200. We denote naive pre-training and fine-tuning as PT+FT and the view-invariant learning as VI-(PT/FT).

Evaluation: We evaluate the performance in the target domain with five experiments with different random seeds. We report the average scores of these two metrics: **dvc_eval** [168] and **SODA** [100]. The dvc_eval metric computes the average precision of the matched pairs between the prediction and the ground truth, namely BLEU4 (B4) [262], METEOR (M) [19], and CIDEr (C) [345]. The SODA metric considers the storytelling quality for an entire video, w.r.t. the order of captions and their redundancy. Following [246], we show METEOR (M), CIDEr (C), and temporal Intersection-over-Union (tIoU) as the SODA scores.

	Method	Input	dvc_eval			SODA		
			B4	M	C	M	C	tIoU
	Source only	V	0.00	0.77	3.6	0.89	1.47	17.9
Baselines	PT+FT	V	1.54	7.03	38.1	7.03	25.2	50.5
		VC	1.97	8.20	46.3	8.04	32.3	55.0
		VC+HO	1.68	8.91	52.5	8.91	37.3	<u>59.0</u>
	+ MMD [338]	VC+HO	1.74	8.86	50.9	8.86	37.5	58.8
	+ DANN [102]	VC+HO	2.05	9.01	53.1	8.97	39.1	58.6
Ours	VI-PT + FT	VC+HO	<u>2.06</u>	**9.44**	<u>55.2</u>	<u>9.02</u>	<u>39.5</u>	56.0
	PT + VI-FT	VC+HO	1.77	8.89	53.0	8.91	37.2	**59.1**
	VI-PT + VI-FT	VC+HO	**2.66**	<u>9.19</u>	**59.0**	**9.27**	**45.2**	58.1

Table 7.3: **Quantitative results in transfer learning from YouCook2 (YC2) to EgoYC2.** We run pre-training (PT) and fine-tuning (FT) with or without the view-invariant (VI) learning. We also compare various input feature types: raw videos (V), cropped videos (VC), and that with hand-object features (VC+HO).

7.5.2 Results

We report the results of the transfer learning task quantitatively and qualitatively and show the ablation results.

Results of transfer learning: Tab. 7.3 shows performance comparisons in the transfer learning setting. The source only shows the zero-shot generalization ability of a trained model on YC2 to EgoYC2. The significant view gaps prevent its generalization to egocentric videos. Our pre-training approach (with view-invariant learning) boosts the performance, suggesting that pre-training on the web video benefits video captioning for egocentric videos.

We compare our gradual adaptation with standard domain adaptation methods aligning two-domain features without assuming an intermediate domain, MMD [338] and DANN [102]. Our methods, VI-PT + FT/VI-FT, exhibit better captioning results than those adaptation methods. This confirms the effectiveness of gradually adapting to the *ego* view with the guidance of the intermediate domain.

VI?		dvc_eval			SODA	
PT	FT	B4	M	C	M	C
-	-	0.98	10.00	50.6	12.62	40.4
✓	-	1.62	**10.35**	**55.5**	**13.57**	47.9
-	✓	1.07	10.13	51.7	12.70	40.9
✓	✓	**1.68**	10.20	54.1	13.36	**48.7**

Table 7.4: **Analysis of captioning performance with GT proposals.** We evaluate our comparison models in Tab. 7.3 given ground-truth (GT) time segments. We use the VC+HO feature as the input. "VI" indicates our proposed method of view-invariant learning introduced in Sec. 7.3.2.

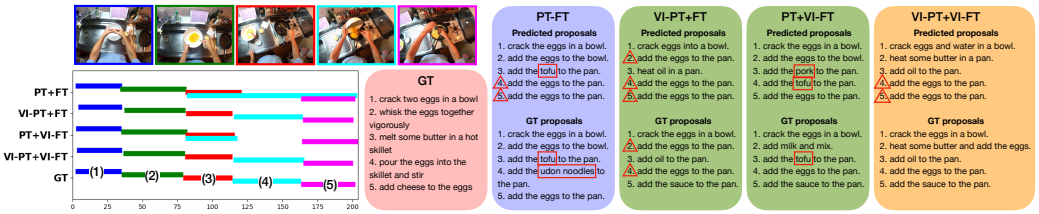


Figure 7.5: **Qualitative results** (recipe: scrambled eggs). We show generated captions given time segment proposals from prediction or ground-truth. We compare our ablation models: view-invariant (VI) pre-training (PT) and/or view-invariant (VI) fine-tuning (FT). The marks ▫ and △ indicate failure cases for irrelevant ingredients and duplicate captions.

Ablation study of egocentric video representations: We validate different representations for egocentric videos in Tab. 7.3, namely raw videos (V), cropped videos (VC), and the VC with hand-object features (VC+HO). We set PT+FT as a base training setting. First, a simple cropping technique with hand tracking (VC) is effective, exhibiting a 21.5 % improvement in the CIDEr score over the V input. This suggests that tracking the hand region reduces the effect of complex scene changes and makes the task more tractable. With the hand-object features (VC+HO), we observe the gain of the CIDEr score by 13.4 % over the VC input and achieve better results.

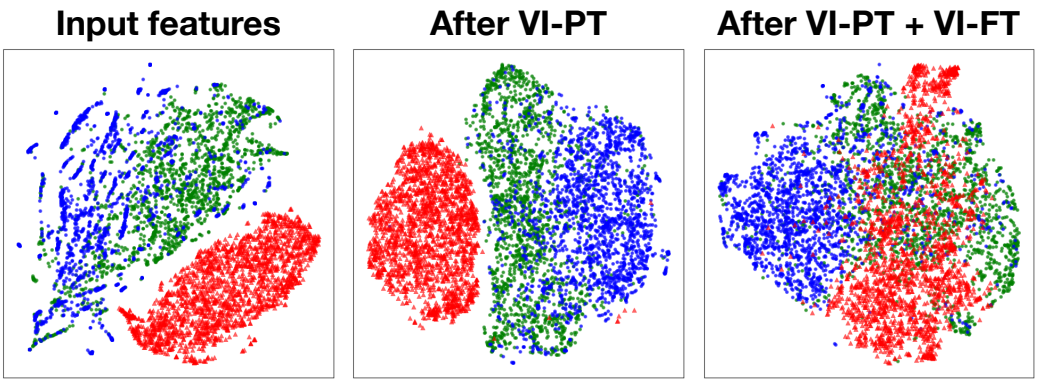


Figure 7.6: **Visualization of feature distribution** (●: *exo*, ●: *ego-like*, ▲: *ego*). We visualize the source and target features encoded in each training stage with t-SNE [211]. Left: initial features generated by the encoder ϕ, Middle: after view-invariant pre-training (VI-PT) on the source data, Right: after view-invariant fine-tuning (VI-PT + VI-FT) on both datasets.

Analysis of captioning capability: Tab. 7.4 shows captioning results given GT time segments without the need for the segment prediction, enabling evaluation of pure captioning ability. Similarly to Tab. 7.3, models based on view-invariant pre-training (VI-PT + FT and full method) exhibit higher captioning performance. Since the given time segments will not overlap, the generated captions have fewer duplicate sentences (see Fig. 7.5).

Qualitative results: Fig. 7.5 shows qualitative results of our ablation models with generated captions and predicted time segments. For time segmentation, while PT+FT and PT + VI-FT generate overlapped segments and overlook some segments, VI-PT + FT and our full method correct these localization errors. In generated captions, we observe several failure patterns: appearing unrelated ingredients and duplicate captions. The captions of PT+FT and PT + VI-FT include the unrelated ingredients (*e.g.*, "tofu", "pork", and "udon noodles"). We also find repeated sentences (triangles) in the models without the view-invariant fine-tuning (*i.e.*, PT+FT and VI-PT + FT). These observations suggest that the view-invariant pre-training reduces the mixing of unrelated ingredients, and the later view-invariant fine-tuning helps produce fewer repeated sentences.

Visualization of feature distribution: To see the transition of the view distribution, we visualize feature distribution for the view classes in Fig. 7.6. The visu-

alization is a way to inspect how well the domain gap is mitigated in the feature space [57, 102, 161, 238], which indicates that the proximity between the points of different domains represents the model's response to the gap. This confirms that our adaptation method aligns *exo* and *ego-like* views on the source data first (after VI-PT) and then aligns all three views (after VI-PT + VI-FT). Interestingly, our method enables aligning visually similar domains with large overlaps between *ego-like* (green) and *ego* (red) views.

Figure 7.7: **Web user interface for our recording.** Top left: Instruction of recording, Top right: Step description with the focus on the current step, Bottom left: Reference video from YouCook2 [407], Bottom right: Necessary ingredients extracted from captions.

7.6 Additional Results and Details

7.6.1 Dataset details

Video recording: We ask 44 participants to record cooking activities in their own home kitchens using a head-mounted GoPro camera. The cooking recipes are adopted from YouCook2 (YC2) [407] captions with 2,000 recipes consisting of 82 classes of recipes (*e.g.*, "BLT" is a class and multiple recipes belong to the class). Each participant chooses five recipes at will so that selected classes do not overlap and then prepares the meal by following the step descriptions written in the recipe. In total, we collect 226 videos totaling 43 hours. We also received approval for this activity data collection from an Institutional Review Board and obtained consent from participants who joined this recording.

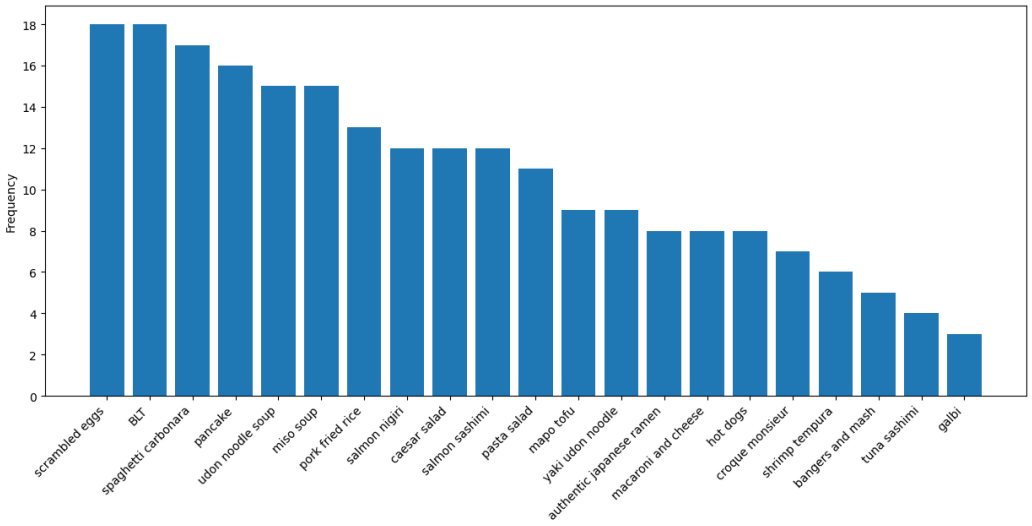


Figure 7.8: **Recipe distribution in EgoYC2**

Fig. 7.7 indicates our web application used for our video recording, displaying the instruction of video collection, step descriptions, reference videos from YouCook2, and necessary ingredients extracted from annotated captions. This Web interface helps participants prepare ingredients and check how to cook from the reference videos prior to recording. During recording, the highlighted step description is shown to indicate their current step and changes by manipulating the button below. The AR markers are displayed on the screen in the transition of steps, which are used to annotate temporal segments.

To maintain the coherency of captured activities, we instruct the participants to remember the recipes beforehand, which allows them to move to the next step smoothly in the actual recording. Even though they halted midway through the recording to remember the step procedure, we treat it as acceptable behavior as it is likely to refer to the recipe on their tablets in real-life cooking.

Transfer learning setup: We use YC2 and EgoYC2 as the source and target data, respectively. We split the EgoYC2 dataset into train and evaluation sets with 151 (964) and 75 (511) videos (step descriptions), respectively. To align both datasets, we re-split the YC2 dataset according to the EgoYC2's split, where train and evaluation sets have 1,716 (13,324) and 75 (511) videos (step descriptions), respectively. The evaluation sets correspond to each other, and all the YC2 data that are not re-recorded in this work are included in the training set.

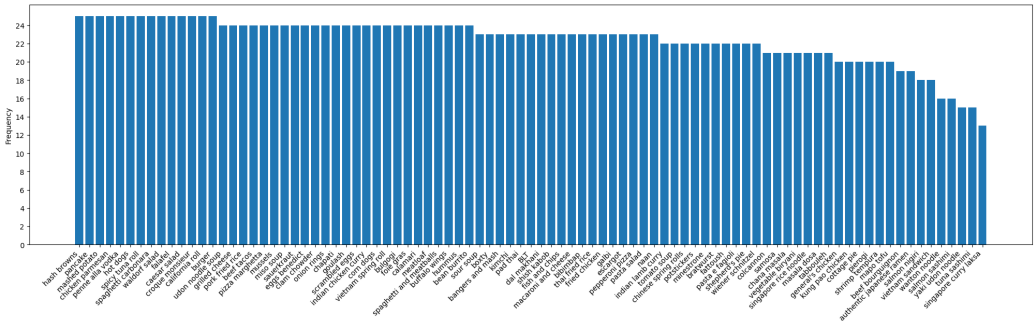


Figure 7.9: **Recipe distribution in YouCook2 [407]**

Recipe class distribution: Figs. 7.8 and 7.9 show the distribution of recipe classes for EgoYC2 and YC2. We collect 21 recipe classes out of 89 classes in YouCook2. The collected recipe list of EgoYC2 is as follows: *BLT*, *authentic japanese ramen*, *bangers and mash*, *caesar salad*, *croque monsieur*, *galbi*, *hot dogs*, *macaroni and cheese*, *mapo tofu*, *miso soup*, *pancake*, *pasta salad*, *pork fried rice*, *salmon nigiri*, *salmon sashimi*, *scrambled eggs*, *shrimp tempura*, *spaghetti carbonara*, *tuna sashimi*, *udon noodle soup*, *yaki udon noodle*.

7.6.2 Additional implementation details

The architectures of the feature converter F and the view classifier C follow a two-layer one-dimensional CNN and a three-layer MLP, respectively. The video features are represented as 2,048-dimensional feature vectors for an input image. We use PDVC [350] as a baseline for dense video captioning. The PDVC uses a two-layer deformable transformer with a hidden size of 512 in the attention layers and 2,048 in the feed-forward layers. The number of event queries is set to 100 and the mini-batch size is set to 1. We use the Adam [163] optimizer with an initial learning rate of 1e-5 for the feature converter and PDVC, and 1e-4 for the view classifier. While we validate various input types for the target egocentric videos, we use the original video features generated by TSN [351] on YouCook2.

7.6.3 Additional results

Results with egocentric data only: Tab. 7.5 shows the results of scratch training on EgoYC2 only. This demonstrates consistent improvement with hand-object

Input	dvc_eval			SODA		
	B4	M	C	M	C	tIoU
V	0.01	3.11	12.3	3.60	5.9	30.7
VC	**0.10**	5.60	22.2	5.62	12.6	41.5
VC+HO	**0.10**	**7.34**	**29.6**	**7.04**	**17.9**	**51.4**

Table 7.5: **Quantitative results in scratch training on EgoYC2.** We train models from scratch in EgoYC2 with various input feature types: raw videos (V), cropped videos (VC), and those with features of an object in hand (VC + HO).

encoding similar to the transfer setup (Tab. 3 in the main paper). With paired videos of YC2 (Rows 2-4 in Tab. 3 in the main paper), we observe significant gains over scratch performance, which confirms the effectiveness of transfer learning in limited data regimes for egocentric video captioning.

Analysis of hyperparameter settings: We set the hyperparameter of view-invariant learning (λ_{adv}) by observing the source performance of the view-invariant pre-training (VI-PT). We use the sum of two METEOR metrics (sum_METEOR) for the model selection during the pre-training. We choose the hyperparameter with the highest sum_METEOR value and set λ_{adv} as 0.1 consistently for the fine-tuning in the target domain.

We also evaluate performance in the pre-training and fine-tuning stages, according to different hyperparameters in Tab. 7.6. Pre-training with $\lambda_{\text{adv}} = 0.01, 0.1$ achieves relatively high performance, while fine-tuning with $\lambda_{\text{adv}} = 0.01, 1$ worsens performance than the PT+FT baseline (top row). When adding the view-invariant technique to both the pre-training and fine-tuning, we observe an improvement of captioning ability with $\lambda_{\text{adv}} = 0.01, 0.1$, as they are adapted from the pre-training models where the view-invariant learning performs well. Based on this study, our setting of $\lambda_{\text{adv}} = 0.1$ chosen from the source pre-training performs stably in the target domain with both the pre-training and fine-tuning stages.

Hand-object segmentation results: We propose a segmentation refinement scheme based on two segmentation models: EgoHOS [400] and SAM [165]. We show the segmentation results for each method in Fig. 7.10. The EgoHOS inference (left) often has noisy results (*e.g.*, undersegmetation on the top row and incorrect lo-

VI? PT	VI? FT	λ_{adv}	dvc_eval B4	dvc_eval M	dvc_eval C	SODA M	SODA C	SODA tIoU
		0	1.68	8.91	52.5	8.91	37.3	59.0
✓		0.01	**2.20**	**9.45**	52.4	8.99	**39.9**	55.0
✓		0.1	2.06	9.44	**55.2**	**9.02**	39.5	**56.0**
✓		1	1.70	9.29	50.5	8.75	36.4	55.1
	✓	0.01	1.47	8.75	49.8	8.72	35.8	58.8
	✓	0.1	**1.77**	**8.89**	**53.0**	**8.91**	**37.2**	59.1
	✓	1	1.50	8.67	49.4	8.67	35.6	**59.5**
✓	✓	0.01	2.46	**9.60**	53.1	8.99	39.3	55.4
✓	✓	0.1	**2.66**	9.19	**59.0**	**9.27**	**45.2**	**58.1**
✓	✓	1	1.58	9.30	49.7	8.67	35.3	54.9

Table 7.6: **Analysis of hyperparameter settings.** We validate different hyperparameters for the view-invariant learning (λ_{adv}) and show the performance on the target dataset.

calization of long and narrow objects on the middle row). EgoHOS suffers from generalizing to novel real-life environments where diverse object types and shapes could be present. The SAM inference (middle) can segment any kind of object with higher generalization. Our refinement (right) computes the overlap between the two results and outputs the most overlapped segments from the SAM predictions. This enables us to obtain further refined results even in crowded cooking environments (*e.g.*, middle row).

7.6.4 Discussions

Scripted *vs*. unscripted: Scripted and unscripted captures each have pros and cons concerning data realism and annotation quality. While unscripted videos, such as Ego4D [111] and EPIC-KITCHENS [77], reflect actual activities, these videos could include ambiguity in captions from human annotators, affecting the consistency of caption content and granularity. Such inconsistency complicates

Figure 7.10: **Our hand-object segmentation refinement.** Each panel shows segmentation results of EgoHOS [400] (left), SAM [165] (middle), and our refined scheme (right), respectively. Since we don't use hand identity information (right/left), we show merged hand masks compared to the results of EgoHOS.

cross-domain evaluation. Our scripted approach not only aligned the content and granularity between datasets, but also instructed participants to maintain action coherency in Sec. 7.6.1, enabling natural step transitions in captured videos.

Unsupervised methods: Zero-shot generalization and unsupervised adaptation remain challenging in video captioning, as evidenced by the source-only results shown in Tab. 3 of the main paper. Our benchmark provides supervised baselines and evaluations on egocentric videos, setting the stage for future studies to develop unsupervised methods.

Overcoming recipe class gap: As shown in Sec. 7.6.1, the recipe class distribution is not perfectly aligned between YC2 and EgoYC2. In addition to focusing on the view gap addressed in the main paper, resolving category shift [44, 210, 371], the gap in the output (label) space, will be an important future challenge.

Comparison with Ego-Exo4D: We provide the comparison with a recently released Ego-Exo4D dataset [112], featuring synchronized egocentric and exocentric videos with textual annotations. In capture setups, the work follows the strong assumption of time-synchronized and calibrated scenarios, while our captures between YC2 and EgoYC2 are based on a relaxed assumption; they are not synchronized and not captured in the same environment. Regarding its text annotations [1],

[1] https://docs.ego-exo4d-data.org/annotations/atomic_descriptions/

the knowledge of the coherency between steps is not explicitly modeled, as each description is instructed to be annotated independently. In contrast, our procedural captions are intended to model the necessary steps to accomplish a target task, which inherently includes inter-step relationships in the captions.

7.7 Conclusion

We present a novel benchmark for cross-view knowledge transfer of dense video captioning from exocentric to egocentric views, together with a new dataset **EgoYC2**. We collect 11.3% of YouCook2 videos from an egocentric view with aligned captions, enabling transfer learning between both datasets. Our proposed view-invariant learning based on adversarial training succeeds in the pre-training and fine-tuning stages while resolving the mixed source views and the moving target view. We validate our proposed method in the cross-view transfer task with quantitative and qualitative analysis. This benchmark will promote further studies of transfer learning across the two views and modeling to describe egocentric activities in natural language.

Chapter 8

Conclusions and Future Work

This chapter provides a comprehensive summary of the dissertation's key findings and discussions. It also outlines future directions based on the analysis of limitations and insights gained from the presented works.

8.1 Summary

This dissertation underscores the ubiquitous role of human hands, which signal human action, expression, and intent. The centrality of hands shapes our interaction with the physical and virtual worlds, enabling advanced human-centric vision systems and applications. While computer vision has progressed in estimating hand states, persistent challenges remain: (i) precisely tracking hands during complicated, fine-grained contact scenarios, (ii) robustly modeling to bridge the domain gap between studio-based training data and in-the-wild testing conditions, and (iii) effectively linking low-level geometric understanding with high-level semantic comprehension of actions and intentions.

This dissertation addresses these limitations by pursuing the overarching goal of **precise tracking and interpretation of fine-grained hand interactions from real-world visual data**. To achieve this ambitious objective, the research is systematically built upon three fundamental pillars: establishing *data foundation*, developing *robust modeling for fine details*, and effectively *connecting geometry and semantics*.

The dissertation's contributions are detailed across the following chapters.

Chapter 2 provided a comprehensive study of the state-of-the-art in 3D hand capture, annotation, and learning methods. This chapter systematically reviews existing techniques and highlights key challenges, thereby offering insights into potential research directions. Chapter 3 focused on advancing egocentric hand pose estimation under complex object interactions. Through benchmark construction and extensive method analysis, this chapter compiled state-of-the-art and practical solutions for the persistent challenges of hand-object contact. Chapter 4 further presented a large-scale benchmark for hand self-contact. This chapter presented generative pose modeling with denoising diffusion networks that effectively addresses the unique self-occlusions and ambiguities prevalent in self-contact scenarios. Chapter 5 directly contributed by presenting a novel approach for 3D hand pose pre-training from diverse, in-the-wild images. This proposed framework aims to build more generalizable and robust hand models capable of performing well across varied conditions. Chapter 6 continued to advance domain adaptation of hand state estimation in under-constrained scenarios. This work detailed methods to bridge performance gaps across different recording environments and camera settings, based on self-training techniques. Chapter 7 addressed the critical connection between geometry and semantics by introducing a novel task for dense video captioning derived from hand-object tracklets. The proposed method demonstrates a pioneering approach that leverages precise geometric information for rich semantic interpretation of hand activities using natural language descriptions.

Collectively, the research presented across these chapters establishes a consistent and integrated framework that significantly advances the visual understanding of hands under complex interactions. By pushing the boundaries in data acquisition, robust estimation, and semantic interpretation, this dissertation lays foundational groundwork for the next generation of intelligent and intuitive human-AI systems.

8.2 Future Work

Despite the advancements presented in this dissertation, several compelling avenues for future research remain, building upon the established framework and addressing inherent limitations.

8.2.1 Expanding data acquisition, sensors, and captured scenarios

While our work has addressed critical gaps in hand interaction datasets, particularly concerning hand-object and self-contact, opportunities exist to broaden this foundation even further. Future efforts should prioritize *more diversified capture settings*, encompassing a wider range of object types, environmental conditions (*e.g.*, extreme lighting, cluttered backgrounds, dynamic scenes), different camera sensors (*e.g.*, static cameras, synchronized AR/VR headsets), and hand appearances (*e.g.*, varied skin tones, wearing gloves [18, 103]). Crucially, this expansion should also encompass capturing hand interactions in 3D with *richer semantic annotations*, *e.g.*, natural language captions for 3D hand captures [99], long-range video captures with narrations describing the procedure [77], and multi-modal semantic context [73]. Furthermore, exploring extended *capture modalities* beyond standard RGB, such as event cameras for high-speed motion [152, 294], thermal cameras for robust sensing in challenging conditions [80], or integrating physical force and haptic feedback [404] (*e.g.*, through instrumented gloves), could provide complementary information for a more comprehensive understanding.

8.2.2 Modeling for temporal context, human modalities, and real-time inference

The pursuit of robust and reliable hand understanding in unconstrained real-world scenarios demands continuous advancement in modeling techniques. A promising target is to achieve *temporally consistent tracking* of 3D hand states in the wild [389, 399], pushing performance beyond frame-wise predictions into temporal or world-aligned expansion. Future research should also investigate the relationship of hand states to *other human modalities* [384] such as body pose [243], head orientation, and eye gaze, allowing for a more holistic and context-aware understanding of human behavior. Furthermore, optimizing these advanced models for *real-time, on-device deployment* [61] remains a crucial step towards ubiquitous application in areas like mobile AR/VR devices and embedded robotic systems. This approach necessitates further investigation into *architecture-specific inductive biases* for optimal efficiency and accuracy, such as CNNs [131, 310], Trans-

formers [344] (including ViTs [83, 203]), and Mamba architectures [81, 114, 416].

8.2.3 Leveraging generative, foundation, and world models

The rapid advancements in large-scale machine learning models offer significant potential for enhancing hand understanding. Future work can further investigate generative approaches (GenAI), including leveraging image or video generation models for providing complementary data sources [54, 139, 291], learning robust generative priors for 3D pose and shape [74], or enabling direct 3D hand generation from sparse image inputs to resolve ambiguities [199, 269]. This also includes strategically leveraging recent foundation models designed for general vision tasks, such as segmentation (SAMs [165, 282]), depth prediction [378, 379], point tracking [69], and human estimation (*e.g.*, Sapiens [158]). Furthermore, world models [21, 22, 219] offer predictive dynamics and representations learned through everyday human activity videos. This can enable more intelligent and anticipatory understanding of the world, where the system not only perceives current hand states but also predicts future outcomes of interactions and understands the physical consequences of actions within a learned environment.

8.2.4 Integrate with common-sense knowledge and reasoning

A critical future direction is to imbue hand understanding systems with common-sense knowledge and reasoning capabilities. This is crucial for interpreting human behavior beyond literal movements, enabling models to understand atypical interactions, infer underlying motivations, and even predict the consequences of actions within various contexts. This includes developing the ability to infer an object's affordances [124], the potential actions or uses an object offers, which is vital for comprehending the rationale behind hand-object interactions. This can be achieved by leveraging the vast knowledge embedded within *Large Language Models (LLMs)* or their multi-modal extensions (MLLMs) [8, 261], which can provide a rich source of common-sense facts and reasoning abilities. By grounding hand semantics in this broader human knowledge, systems can achieve sharper interpretations, moving closer to understanding why an action is performed, rather than just what is being done. This integration can also facilitate the disambiguation of geometrically similar actions based on context and implied intent.

8.2.5 Towards social and collaborative interactions

Moving beyond individual hand understanding, the next frontier involves comprehending complex multi-person and social hand interactions. This includes analyzing collaborative hand movements in cooperative work scenarios [352], understanding communicative gestures in conversational settings [244], or interpreting strategies in social games like card games [16]. Compared to single-person settings, multi-person scenarios introduce the need to model not only spatial coordination but also temporal alignment of gestures with co-occurring speech, requiring models to infer joint intent, shared goals, and the dynamic interplay [155, 178] between multiple agents' hands and verbal cues. These joint modeling and multimodal analysis offer richer insights into human social behavior and enable more sophisticated human-robot collaboration.

8.2.6 Physics-based simulation

Recent advances highlight the potential of physics-based simulation (*e.g.*, MuJoCo [332], Isaac Gym [213], Genesis [12]) for both data generation and model development. These simulators can serve as an invaluable source for vast and precisely annotated synthetic data, explicitly covering difficult and rare contact scenarios that are challenging to capture and annotate in the real world (*e.g.*, specific object deformations, complex grasps, high-speed collisions). These are further useful for simulating the appearance, texture, and lighting of hands [67, 404]. Beyond data generation, physics simulations can validate the physical plausibility of 3D hand predictions against physical law [127, 339] or biomechanics, including an anatomically realistic skeleton [157, 364] and muscle structures [188]. Furthermore, physics-based environments can facilitate the learning of dexterous manipulation policies in simulation [132, 276, 277], which can then be transferred to real robotic systems. This physics-based approach can bridge the gap between perception and action, paving the way for more physically grounded intelligent systems.

Bibliography

[1] A. Abdullah, J. Kolkmeier, V. Lo, and M. Neff. Videoconference and embodied VR: Communication patterns across task and medium. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):453:1–453:29, 2021.

[2] K. Abou Zeid. JointTransformer: Winner of the HANDS'2023 ARCTIC Challenge @ ICCV. https://github.com/kabouzeid/JointTransformer, 2023.

[3] I . Akhter and M . J . Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455, 2015.

[4] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4575–4583, 2016.

[5] F. Alemuda and F.J. Lin. Gesture-based control in a smart home environment. In *IEEE International Conference on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social Computing and IEEE Smart Data*, pages 784–791, 2017.

[6] B. Alsharif, E. Alalwany, A. Ibrahim, I. Mahgoub, and M. Ilyas. Real-time american sign language interpretation using deep learning and keypoint tracking. *Sensors*, 25(7):2138, 2025.

[7] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014.

[8] R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, D. Silver, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen, E. Pitler, T. P. Lillicrap, A. Lazaridou, O. Firat, J. Molloy, M. Isard, P. R. Barham, T. Hennigan, B. Lee, F. Viola, M. Reynolds, Y. Xu, R. Doherty, E. Collins, C. Meyer, E. Rutherford, E. Moreira, K. Ayoub, M. Goel, J. Krawczyk, C. Du, E. Chi, H.-T. Cheng, E. Ni, P. Shah, P. Kane, B. Chan, M. Faruqui, A. Severyn, H. S. Rae, H. Lu,

L. Sifre, M. Maggioni, F. Alcober, D. Garrette, M. Barnes, S. Thakoor, J. Austin, G. Barth-Maron, W. Wong, R. Joshi, R. Chaabouni, D. Fatiha, A. Ahuja, G. S. Tomar, E. Senter, M. Chadwick, I. Kornakov, N. Attaluri, I. Iturrate, R. Chang, A. Recasens, B. Caine, A. Pritzel, F. Pavetic, F. Pardo, A. Gergely, J. Frye, V. Ramasesh, D. Horgan, K. Badola, N. Kassner, S. Roy, E. Dyer, V. C. Campos, A. Tomala, Y. Tang, D. E. Badawy, E. White, B. Mustafa, O. Lang, A. Jindal, A. Balakrishna, R. Baruch, M. Bauzá, M. Blokzijl, S. Bohez, K. Bousmalis, A. Brohan, T. Buschmann, A. Byravan, S. Cabi, K. Caluwaerts, F. Carnevale, M. Cassin, T. von Glehn, A. Goldin, L. Gonzalez, A. G. Arenas, P. C. Humphreys, A. Hung, R. Ives, J. Keeling, M. Khalman, M. Krikun, J. Landon, K. Lenc, D. Lepikhin, J. Lhotka, R. McIlroy, H. Michalewski, S. Modi, A. Muldal, N. Savinov, N. Schucher, E. Sezener, S. Shakeri, P. Shyam, T. Sottiaux, S. Spencer, J. Sygnowski, D. Teplyashin, G. Thornton, G. Tucker, A. White, N. Wong, Y. Wu, L. Yagati, Z. Yang, C. Yan, and R. Zhu. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023.

[9] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307, 2016.

[10] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

[11] A. Armagan, G. G.-Hernando, S. Baek, S. Hampali, M. Rad, Z. Zhang, S. Xie, M. Chen, B. Zhang, F. Xiong, Y. Xiao, Z. Cao, J. Yuan, P. Ren, W. Huang, H. Sun, M. Hrúz, J. Kanis, Z. Krňoul, Q. Wan, S. Li, L. Yang, D. Lee, A. Yao, W. Zhou, S. Mei, Y. Liu, A. Spurr, U. Iqbal, P. Molchanov, P. Weinzaepfel, R. Brégier, G. Rogez, V. Lepetit, and T.-K. Kim. Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3D hand pose estimation under hand-object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–101, 2020.

[12] Genesis Authors. Genesis: A universal and generative physics engine for

robotics and beyond, December 2024.

[13] O . Avrahami, O . Fried, and D . Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (ToG)*, 42(4):149:1–149:11, 2023.

[14] S. Baek, K. I. Kim, and T.-K. Kim. 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)*, pages 6120–6130, 2020.

[15] L. Ballan, A. Taneja, J. Gall, L. V. Gool, and M. Pollefeys. Motion capture of hands in action using discriminative salient points. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 7577, pages 640–653, 2012.

[16] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1949–1957, 2015.

[17] A. Bandini and J. Zariffa. Analysis of the hands in egocentric vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–1, 2020.

[18] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol, R. Newcombe, R. Wang, J. J. Engel, and T. Hodan. Hot3d: An egocentric dataset for 3d hand and object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[19] S. Banerjee and A. Lavie. METEOR: an automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

[20] S. Bansal, C. Arora, and C. V. Jawahar. My view is the best view: Procedure learning from egocentric videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 657–675, 2022.

[21] A. Bar, Z. Gan, F. Yu, Z. Yang, C. Zhou, P. C. Tursun, T. Darrell, X. Li, Y. LeCun, J. J. Lim, X. Wang, and H. Fang. GEM: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), 2025.

[22] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[23] K. Bartol, D. Bojanić, T. Petković, and T. Pribanić. Generalizable human pose triangulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11018–11027, 2022.

[24] G. Benitez-Garcia, L. Prudente-Tixteco, L. C. Castro-Madrid, R. Toscano-Medina, J. Olivares-Mercado, G. Sanchez-Perez, and L. J. G. Villalba. Improving real-time hand gesture recognition with semantic segmentation. *IEEE Sensors*, 21(2), 2021.

[25] G. Berton, C. Masone, and B. Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4878–4888, 2022.

[26] A. Bewley, Z. Ge, L. Ott, F. T. Ramos, and B. Upcroft. Simple online and realtime tracking. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, 2016.

[27] M. Bianchi, P. Salaris, and A. Bicchi. Synergy-based hand pose sensing: Optimal glove design. *The International Journal of Robotics Research (IJRR)*, 32(4):407–424, 2013.

[28] M . Binkowski, D . J . Sutherland, M . Arbel, and A . Gretton. Demystifying MMD GANs. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018.

[29] E. Bkheet, A.-L. D'Angelo, A. Goldbraikh, and S. Laufer. Using hand pose estimation to automate open surgery training feedback. *International Journal of Computer Assisted Radiology and Surgery*, 18(7):1279–1285, 2023.

[30] A. Blum and T. M. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the ACM Annual Conference on Computational Learning Theory (COLT)*, pages 92–100, 1998.

[31] A. Boukhayma, R. A. Bem, and P. H. S. Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10843–10852,

2019.

[32] A. Boukhayma, R. de Bem, and P. H. S. Torr. 3D hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10843–10852, 2019.

[33] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 361–378, 2020.

[34] J. Bronowski. *The Ascent of Man*. BBC Books, 1973.

[35] A. Brunzini, M. Ciccarelli, M. Sartini, and M. Germani. Evaluation of vision-based hand tool tracking methods for quality assessment and training in human-centered industry 4.0. *Applied Sciences*, 12(4):1796, 2022.

[36] G. Buckingham. Hand tracking for immersive virtual reality: Opportunities and challenges. *Frontiers in Virtual Reality*, 2, 2021.

[37] M. Cai, E. Lu, and Y. Sato. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14380–14389, 2020.

[38] M. Cai, M. Luo, X. Zhong, and H. Chen. Uncertainty-aware model adaptation for unsupervised cross-domain object detection. *CoRR*, abs/2108.12612, 2021.

[39] Q. Cai, Y. Pan, C.-W. Ngo, X. Tian, L. Duan, and T. Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11457–11466, 2019.

[40] Y. Cai, L. Ge, J. Cai, and J. Yuan. Weakly-supervised 3D hand pose estimation from monocular RGB images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 678–694, 2018.

[41] Z . Cai, W . Yin, A . Zeng, C . Wei, Q . Sun, W . Yanjun, H . E . Pang, H . Mei, M . Zhang, L . Zhang, C . C . Loy, L . Yang, and Z . Liu. Smpler-x: Scaling up expressive human pose and shape estimation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[42] B. Çalli, A. Walsman, A. Singh, S. S. Srinivasa, P. Abbeel, and A. M. Dol-

lar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics Automation Magazine*, 22(3):36–52, 2015.

[43] J. Cao, H. Tang, H. Fang, X. Shen, Y.-W. Tai, and C. Lu. Cross-domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9497–9506, 2019.

[44] Z. Cao, L. Ma, M. Long, and J. Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–155, 2018.

[45] Z. Cao, I. Radosavovic, A. Kanazawa, and J. Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12417–12426, 2021.

[46] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 9912–9924, 2020.

[47] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021.

[48] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q.-X. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015.

[49] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, J. Kautz, and D. Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9044–9053, 2021.

[50] D. Chatterjee, F. Sener, S. Ma, and A. Yao. Opening the vocabulary of egocentric actions. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[51] T. Chatzis, A. Stergioulas, D. Konstantinidis, K. Dimitropoulos, and P.

Daras. A comprehensive study on deep learning-based 3d hand pose estimation methods. *Applied Sciences*, 10:6850, 09 2020.

[52] C.-H. Chen, A. Tyagi, A. Agrawal, D. Drover, R. MV, S. Stojanov, and J. M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5714–5724, 2019.

[53] H. Chen, Y. Wang, B. Lagadec, A. Dantcheva, and F. Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2004–2013, 2021.

[54] K. Chen, C. Min, L. Zhang, S. Hampali, C. Keskin, and S. Sridhar. Foundhand: Large-scale domain-specific learning for controllable hand image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[55] L. Chen, S.-Y. Lin, Y. Xie, Y.-Y. Lin, and X. Xie. MVHM: A large-scale multi-view hand mesh benchmark for accurate 3d hand pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 836–845, 2021.

[56] M. Chen, K. Q. Weinberger, and J. Blitzer. Co-training for domain adaptation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 2456–2464, 2011.

[57] M.-H. Chen, Z. Kira, G. Alregib, J. Yoo, R. Chen, and J. Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6320–6329, 2019.

[58] S. Chen, X. Nie, D. Fan, D. Zhang, V. Bhat, and R. Hamid. Shot contrastive self-supervised learning for scene boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9796–9805, 2021.

[59] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, pages 1597–1607, 2020.

[60] X. Chen and K. He. Exploring simple siamese representation learning. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 15750–15758, 2021.

[61] X. Chen, Y. Liu, Y. Dong, X. Zhang, C. Ma, Y. Xiong, Y. Zhang, and X. Guo. MobRecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20512–20522, 2022.

[62] X. Chen, B. Wang, and H.-Y. Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8683–8693, 2023.

[63] X. Chen, G. Wang, C. Zhang, T.-K. Kim, and X. Ji. Shpr-net: Deep semantic hand pose regression from point clouds. *IEEE Access*, 6:43425–43439, 2018.

[64] Y. Chen, S. K. Dwivedi, M. J. Black, and D. Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17100–17110, 2023.

[65] Y. Chen, Z. Tu, D. Kang, L. Bao, Y. Zhang, X. Zhe, R. Chen, and J. Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10451–10460, 2021.

[66] Z. Chen, S. Chen, C. Schmid, and I. Laptev. gSDF: Geometry-driven signed distance functions for 3D hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12890–12900, 2023.

[67] Z. Chen, G. Moon, K. Guo, C. Cao, S. Pidhorskyi, T. Simon, R. Joshi, Y. Dong, Y. Xu, B. Pires, H. Wen, L. Evans, B. Peng, J. Buffalini, A. Trimble, K. McPhail, M. Schoeller, S. Yu, J. Romero, M. Zollhöfer, Y. Sheikh, Z. Liu, and S. Saito. Urhand: Universal relightable hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 119–129, 2024.

[68] Z. Chen, S. Wang, Y. Sun, and X. Ma. Self-supervised transfer learning for hand mesh recovery from binocular images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages

11606–11614, 2021.

[69] T. Cheng, K. Liao, L. Shi, B. Yang, A. E. M. T. Mohamed, K. G. Derpanis, Y. Li, and Z. Zhang. CoTracker: It is better to track together. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10034–10043, 2024.

[70] J. Choi, G. Sharma, M. Chandraker, and J.-B. Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1706–1715, 2020.

[71] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 539–546, 2005.

[72] S. Christen, M. Kocabas, E. Aksan, J. Hwangbo, J. Song, and O. Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20545–20554, 2022.

[73] P. Chua, C. M. Fang, T. Ohkawa, R. Kushalnagar, S. Nanayakkara, and P. Maes. Emosign: A multimodal dataset for understanding emotions in american sign language. *CoRR*, abs/2505.17090, 2025.

[74] H. Ci, M. Wu, W. Zhu, X. Ma, H. Dong, F. Zhong, and Y. Wang. GFPose: Learning 3d human pose prior with gradient fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4800–4810, 2023.

[75] M. T. Ciocarlie and P. K. Allen. Hand posture subspaces for dexterous robotic grasping. *The International Journal of Robotics Research (IJRR)*, 28(7):851–867, 2009.

[76] E. Corona, A. Pumarola, G. Alenyà, F. Moreno-Noguer, and G. Rogez. GanHand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5030–5040, 2020.

[77] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Rescaling egocentric vision. *International Journal of Computer Vision (IJCV)*, 130(1):33–

55, 2022.

[78] A. Darkhalil, D. Shan, B. Zhu, J. Ma, A. Kar, R. E. L. Higgins, S. Fidler, D. Fouhey, and D. Damen. EPIC-KITCHENS VISOR benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022.

[79] C. Deng, S. Chen, D. Chen, Y. He, and Q. Wu. Sketch, ground, and refine: Top-down dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 234–243, 2021.

[80] F. Ding, Y. Zhu, X. Wen, G. Liu, and C. X. Lu. Thermohands: A benchmark for 3d hand pose estimation from egocentric thermal images. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pages 533–546, 2025.

[81] H. Dong, A. Chharia, W. Gou, F. V. Carrasco, and F. D. Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[82] B. Doosti. Hand pose estimation: A survey. *CoRR*, abs/1903.01013, 2019.

[83] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[84] E. Duran, M. Kocabas, V. Choutas, Z. Fan, and M. J. Black. HMP: Hand motion priors for pose and shape estimation from video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6341–6351, 2024.

[85] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith, C. Peng, C. Sweeney, C. Wilson, D. Barnes, D. DeTone, D. Caruso, D. Valleroy, D. Ginjupalli, D. Frost, E. Miller, E. Mueggler, E. Oleinik, F. Zhang, G. Somasundaram, G. Solaira, H. Lanaras, H. Howard-Jenkins, H. Tang, H.J. Kim, J. Rivera, and J. Luo. Project Aria: A new tool for egocentric multi-modal AI research. *CoRR*, abs/2308.13561, 2023.

[86] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding (CVIU)*, 108(1-2):52–73, 2007.

[87] M. Eswaran, V.S.S.V. Prasad, M. Hymavathi, and M.V.A.R. Bahubalendruni. Augmented reality guided autonomous assembly system: A novel framework for assembly sequence input validations and creation of virtual content for AR instructions development. *Journal of Manufacturing Systems*, 72:104–121, 2024.

[88] Z . Fan, O . Taheri, D . Tzionas, M . Kocabas, M . Kaufmann, M . J . Black, and O . Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12943–12954, 2023.

[89] Z. Fan, T. Ohkawa, L. Yang, N. Lin, Z. Zhou, S. Zhou, J. Liang, Z. Gao, X. Zhang, X. Zhang, F. Li, L. Zheng, F. Lu, K. A. Zeid, B. Leibe, J. On, S. Baek, A. Prakash, S. Gupta, K. He, Y. Sato, O. Hilliges, H. J. Chang, and A. Yao. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15083, pages 428–448, 2024.

[90] Z. Fan, M. Parelli, M. E. Kadoglou, X. Chen, M. Kocabas, M. J. Black, and O. Hilliges. HOLD: Category-agnostic 3D reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 494–504, 2024.

[91] Z. Fan, A. Spurr, M. Kocabas, S. Tang, M. J. Black, and O. Hilliges. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 1–10, 2021.

[92] Q. Feng, K. He, H. Wen, C. Keskin, and Y. Ye. Active learning with pseudo-labels for multi-view 3d pose estimation. *CoRR*, abs/2112.13709, 2021.

[93] M . Fieraru, M . Zanfir, E . Oneata, A . Popa, V . Olaru, and C . Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7212–7221, 2020.

[94] M . Fieraru, M . Zanfir, E . Oneata, A . Popa, V . Olaru, and C . Sminchis-

escu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1343–1351, 2021.

[95] G. French, M. Mackiewicz, and M. H. Fisher. Self-ensembling for visual domain adaptation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[96] H. Fu, M. Gong, C. Wang, K. Batmanghelich, K. Zhang, and D. Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2427–2436, 2019.

[97] Q. Fu, X. Liu, and K. M. Kitani. Sequential decision-making for active object detection from hand. *CoRR*, abs/2110.11524, 2021.

[98] Q. Fu, X. Liu, R. Xu, J. C. Niebles, and K. M. Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23543–23554, 2023.

[99] R. Fu, D. Zhang, A. Jiang, W. Fu, A. Funk, D. Ritchie, and S. Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[100] S. Fujita, T. Hirao, H. Kamigaito, M. Okumura, and M. Nagata. SODA: story oriented dense video captioning evaluation framework. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 517–531, 2020.

[101] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1050—-1059, 2016.

[102] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1180–1189, 2015.

[103] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition (CVPR), pages 409–419, 2018.

[104] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan. 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)*, pages 10833–10842, 2019.

[105] Y. Ge, D. Chen, and H. Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[106] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li. Self-supervising fine-grained region similarities for large-scale image localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 369–386, 2020.

[107] O. Glauser, S. Wu, D. Panozzo, O. Hilliges, and O. Sorkine-Hornung. Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Transactions on Graphics (ToG)*, 38(4):41:1–41:15, 2019.

[108] D. Goudie and A. Galata. 3D hand-object pose estimation from depth with convolutional neural networks. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 406–413, 2017.

[109] M. Goyal, S. Modi, R. Goyal, and S. Gupta. Human hands as probes for interactive object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3283–3293, 2022.

[110] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1471–1481, 2021.

[111] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M.g Xu, E. Zhongcong Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li,

K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. Soo Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, Lo Torresani, M.i Yan, and J. Malik. 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)*, pages 18973–18990, 2022.

[112] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, A. Kumar, V. Baiyya, S. Bansal, B. Boote, E. Byrne, Z. Chavis, J. Chen, F. Cheng, F. Chu, S. Crane, A. Dasgupta, J. Dong, M. Escobar, C. Forigua, A. Gebreselasie, S. Haresh, J. Huang, M. M. Islam, S. Jain, R. Khirodkar, D. Kukreja, K. J. Liang, J. Liu, S. Majumder, Y. Mao, M. Martin, E. Mavroudi, T. Nagarajan, F. Ragusa, S. K. Ramakrishnan, L. Seminara, A. Somayazulu, Y. Song, S. Su, Z. Xue, E. Zhang, J. Zhang, A. Castillo, C. Chen, X. Fu, R. Furuta, C. Gonzalez, P. Gupta, J. Hu, Y. Huang, Y. Huang, W. Khoo, A. Kumar, R. Kuo, S. Lakhavani, M. Liu, M. Luo, Z. Luo, B. Meredith, A. Miller, O. Oguntola, X. Pan, P. Peng, S. Pramanick, M. Ramazanova, F. Ryan, W. Shan, K. Somasundaram, C. Song, A. Southerland, M. Tateno, H. Wang, Y. Wang, T. Yagi, M. Yan, X. Yang, Z. Yu, S. C. Zha, C. Zhao, Z. Zhao, Z. Zhu, J. Zhuo, P. Arbelaez, G. Bertasius, D. Damen, J. Engel, G. M. Farinella, A. Furnari, B. Ghanem, J. Hoffman, C. V. Jawahar, R. Newcombe, H. S. Park, J. M. Rehg, Y. Sato, M. Savva, J. Shi, M. Z. Shou, and M. Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19383–19400, 2024.

[113] J. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 21271–21284, 2020.

[114] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024.

[115] Z. Guo, W. Zhou, M. Wang, L. Li, and H. Li. Handnerf: Neural radiance fields for animatable interacting hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21078–21087, 2023.

[116] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit. Honnotate: A method for 3D annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3196–3206, 2020.

[117] S. Hampali, S. D. Sarkar, M. Rad, and V. Lepetit. 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)*, pages 11080–11090, 2022.

[118] S. Han, B. Liu, R. Cabezas, C. D. Twigg, P. Zhang, J. Petkau, T.-H. Yu, C.-J. Tai, M. Akbay, Z. Wang, A. Nitzan, G. Dong, Y. Ye, L. Tao, C. Wan, and R. Wang. MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG)*, 39(4):87, 2020.

[119] S. Han, P.-C. Wu, Y. Zhang, B. Liu, L. Zhang, Z. Wang, W. Si, P. Zhang, Y. Cai, T. Hodan, R. Cabezas, L. Tran, M. Akbay, T.-H. Yu, C. Keskin, and R. Wang. UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *In Proceedings of the ACM SIGGRAPH Asia Conference*, pages 50:1–50:9, 2022.

[120] A. Handa, K. V. Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox. DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 9164–9170, 2020.

[121] J. A Harrigan. Self-touching as an indicator of underlying affect and language processes. *Social Science & Medicine*, 20(11):1161–1168, 1985.

[122] J. A. Harrigan, K. S. Lucic, D. Kay, A. McLaney, and R. Rosenthal. Effect of expresser role and type of self-touching on observers' perceptions 1. *Journal of Applied Social Psychology*, 21(7):585–609, 1991.

[123] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black. Resolving 3d hu-

man pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019.

[124] M. Hassanin, S. Khan, and M. Tahtali. Visual affordance and function understanding: A survey. *ACM Computing Survey*, 54(3):47:1–47:35, 2021.

[125] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and a. C. Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 568–577, 2020.

[126] Y. Hasson, G. Varol, I. Laptev, and C. Schmid. Towards unconstrained joint hand-object reconstruction from RGB videos. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 659–668, 2021.

[127] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019.

[128] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14141–14152, 2021.

[129] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022.

[130] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

[131] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[132] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-

body teleoperation and learning. In *Proceedings of the Conference on Robot Learning (CoRL)*, pages 1516–1540, 2024.

[133] J. Hein, M. Seibold, F. Bogo, M. Farshad, M. Pollefeys, P. Fürnstahl, and N. Navab. Towards markerless surgical tool and hand pose estimation. *International Journal of Computer Assisted Radiology and Surgery*, 16(5):799–808, 2021.

[134] M . Heusel, H . Ramsauer, T . Unterthiner, B . Nessler, and S . Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017.

[135] G. Hidalgo, Z. Cao, T. Simon, S.-E. Wei, Y. Raaj, H. Joo, and Y. Sheikh. OpenPose. https://github.com/CMU-Perceptual-Computing-Lab/openpose.

[136] E. S. L. Ho, H. Wang, and T. Komura. A multi-resolution approach for adapting close character interaction. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST)*, pages 97–106, 2014.

[137] J . Ho, A . Jain, and P . Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[138] J . Ho and T . Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022.

[139] J. Ho, T. Salimans, A. A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[140] H . Hu, Z . Fan, T . Wu, Y . Xi, S . Lee, G . Pavlakos, and Z . Wang. Expressive gaussian human avatars from monocular RGB video. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[141] C.-H. P. Huang, H. Yi, M. Höschle, M. Safroshkin, T. Alexiadis, S. Polikovsky, D. Scharstein, and M. J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13264–13275, 2022.

[142] C.-M. Huang and B. Mutlu. Anticipatory robot control for efficient human-

robot collaboration. In *Proceedings of the International Conference on Human-Robot Interaction (HRI)*, pages 321–328, 2016.

[143] D. Huang, X. Ji, X. He, J. Sun, T. He, Q. Shuai, W. Ouyang, and X. Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia Conference Proceedings*, 2022.

[144] L. Huang, J. Tan, J. Liu, and J. Yuan. Hand-transformer: Non-autoregressive structured modeling for 3d hand pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12370, pages 17–33, 2020.

[145] W. Huang, P. Ren, J. Wang, Q. Qi, and H. Sun. AWR: Adaptive weighting regression for 3D hand pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 11061–11068, 2020.

[146] Y. Huang, L. Yang, and Y. Sato. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18908–18918, 2023.

[147] J. S. Supancic III, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan. Depth-based hand pose estimation: Methods, data, and challenges. *International Journal Computer Vision (IJCV)*, 126(11):1180–1198, 2018.

[148] U. Iqbal, M. Garbade, and J. Gall. Pose for action - action for pose. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, FG*, pages 438–445, 2017.

[149] U. Iqbal, P. Molchanov, T. M. Breuel, J. Gall, and J. Kautz. Hand pose estimation via latent 2.5D heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 125–143, 2018.

[150] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7718–7727, 2019.

[151] J. Jiang, Y. Ji, X. Wang, Y. Liu, J. Wang, and M. Long. Regressive domain adaptation for unsupervised keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6780–6789, 2021.

[152] J. Jiang, J. Li, B. Zhang, X. Deng, and B. Shi. Evhandpose: Event-based 3d hand pose estimation with sparse supervision. *IEEE Transactions on Pat-*

tern Analysis and Machine Intelligence (TPAMI), 46(9):6416–6430, 2024.

[153] X. Jie, S. Chen, Y. Ren, X. Shi, H. Shen, G. Niu, and X. Zhu. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 1119–1131, 2024.

[154] S . Jin, L . Xu, J . Xu, C . Wang, W . Liu, C . Qian, W . Ouyang, and P . Luo. Whole-body human pose estimation in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 12354, pages 196–214, 2020.

[155] H. Joo, H. Liu, L. Tan, L. Gui, B. C. Nabbe, I. A. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3334–3342, 2015.

[156] S.-H. Kang and J.-H. Han. Video captioning based on both egocentric and exocentric views of robot vision for human-robot interaction. *International Journal of Social Robotics*, 15(4):631–641, 2023.

[157] M. Keller, K. Werling, S. Shin, S. L. Delp, S. Pujades, C. K. Liu, and M. J. Black. From skin to skeleton: Towards biomechanically accurate 3d digital humans. *ACM Transactions on Graphics (ToG)*, 42(6):253:1–253:12, 2023.

[158] R . Khirodkar, T . M . Bagautdinov, J . Martinez, S . Zhaoen, A . James, P . Selednik, S . Anderson, and S . Saito. Sapiens: Foundation for human vision models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15062, pages 206–228, 2024.

[159] R. Khirodkar, A. Bansal, L. Ma, R. Newcombe, M. Vo, and K. Kitani. Ego-humans: An ego-centric 3d multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19807–19819, 2023.

[160] R. Khirodkar, J.-T. Song, J. Cao, Z. Luo, and K. Kitani. Harmony4D: A video dataset for in-the-wild close human interactions. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[161] D. Kim, Y.-H. Tsai, B. Zhuang, X. Yu, S. Sclaroff, K. Saenko, and M. Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on*

Computer Vision (ICCV), pages 13598–13607, 2021.

[162] S. Kim, H. -G. Chi, X. Hu, A. Vegesana, and K. Ramani. First-person view hand segmentation of multi-modal hand activity video dataset. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[163] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[164] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[165] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.

[166] D. Ko, J. Choi, J. Ko, S. Noh, K.-W. On, E.-S. Kim, and H. J. Kim. Video-text representation learning via differentiable weak temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5006–5015, 2022.

[167] T. Kojić, M. Vergari, S. Knuth, M. Warsinke, S. Möller, and J. Voigt-Antons. Influence of gameplay duration, hand tracking, and controller based control methods on UX in VR. In *International Workshop on Immersive Mixed and Virtual Environment Systems*, pages 22–28, 2024.

[168] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 706–715, 2017.

[169] D. Kulon, R. A. Güler, I. Kokkinos, M. M. Bronstein, and S. Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4989–4999, 2020.

[170] A. Kumar, T. Ma, and P. Liang. Understanding self-training for gradual domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5468–5479, 2020.

[171] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys. H2O: two hands manipulating objects for first person interaction recognition. In *Pro-*

ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 10118–10128, 2021.

[172] V.-H. Le and H.-C. Nguyen. A survey on 3d hand skeleton and pose estimation by convolutional neural network. *Advances in Science, Technology and Engineering Systems Journal (ASTES)*, 5(4):144–159, 2020.

[173] J . Lee, S . Saito, G . Nam, M . Sung, and T . Kim. Interhandgen: Two-hand interaction generation via cascaded reverse diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 527–537. IEEE, 2024.

[174] J. Lee, M. Sung, H. Choi, and T.-K. Kim. Im2hands: Learning attentive implicit representation of interacting two-hand shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21169–21178, 2023.

[175] J. Lee, W. Xu, A. Richard, S.-E. Wei, S. Saito, S. Bai, T.-L. Wang, M. Sung, T.-K. Kim, and J. M. Saragih. REWIND: Real-time egocentric whole-body motion diffusion with exemplar-based identity conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[176] K. Lee, A. Shrivastava, and H. Kacorri. Hand-priming in object localization for assistive egocentric vision. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3422–3432, 2020.

[177] K.-S. Lee and M.-C. Jung. Investigation of hand postures in manufacturing industries according to hand and object properties. *International Journal of Industrial Ergonomics*, 46:98–104, 2015.

[178] S. Lee, B. Lai, F. Ryan, B. Boote, and J. M. Rehg. Modeling multimodal social interactions: New challenges and baselines with densely aligned representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14585–14595, 2024.

[179] V. Lepetit. Recent advances in 3d object and hand pose estimation. *CoRR*, abs/2006.05927, 2020.

[180] J . Li, S . Bian, C . Xu, Z . Chen, L . Yang, and C . Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *CoRR*, abs/2304.05690, 2023.

[181] J. Li, S. Bian, Q .Liu, J. Tang, F. Wang, and C. Lu. NIKI: neural inverse

kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12933–12942. IEEE, 2023.

[182] J. Li, C. Cao, G. Schwartz, R. Khirodkar, C. Richardt, T. Simon, Y. Sheikh, and S. Saito. Uravatar: Universal relightable gaussian codec avatars. In *SIGGRAPH Asia*, pages 128:1–128:11, 2024.

[183] L. Li, L. Tian, X. Zhang, Q. Wang, B. Zhang, L. Bo, M. Liu, and C. Chen. Renderih: A large-scale synthetic dataset for 3D interacting hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20338–20348, 2023.

[184] M. Li, L. An, H. Zhang, L. Wu, F. Chen, T. Yu, and Y. Liu. Interacting attention graph for single image two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2751–2760, 2022.

[185] Y.-J. Li, X. Dai, C.-Y. Ma, Y.-C. Liu, K. Chen, B. Wu, Z. He, K. Kitani, and P. Vajda. Cross-domain object detection via adaptive self-training. *CoRR*, abs/2111.13216, 2021.

[186] Y. Li, T. Nagarajan, B. Xiong, and K. Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6943–6953, 2021.

[187] Y. Li, Z. Ye, and J. M. Rehg. Delving into egocentric actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 287–295, 2015.

[188] Y. Li, L. Zhang, Z. Qiu, Y. Jiang, N. Li, Y. Ma, Y. Zhang, L. Xu, and J. Yu. NIMBLE: A non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (ToG)*, 41(4):120:1–120:16, 2022.

[189] Z. Li, S. Shimada, B. Schiele, C. Theobalt, and V. Golyanik. MoCapDeform: Monocular 3d human motion capture in deformable scenes. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 1–11, 2022.

[190] H. Liang, J.G. Yuan, D. Thalmann, and N. Magnenat-Thalmann. AR in hand: Egocentric palm pose tracking and gesture recognition for augmented reality applications. In *Proceedings of the ACM International Con-*

ference on Multimedia (MM), pages 743–744, 2015.

[191] J. Likitlersuang, E. R. Sumitro, T. Cao, R. J. Visée, S. Kalsi-Ryan, and J. Zariffa. Egocentric video: A new tool for capturing hand use of individuals with spinal cord injury at home. *Journal of Neuroengineering and Rehabilitation (JNER)*, 16(1):83, 2019.

[192] J . Lin, A . Zeng, H . Wang, L . Zhang, and Y . Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21159–21168, 2023.

[193] K. Q. Lin, J. Wang, M. Soldan, M. Wray, R. Yan, E. Z. Xu, D. Gao, R.-C. Tu, W. Zhao, W. Kong, C. Cai, H. Wang, D. Damen, B. Ghanem, W. Liu, and M. Z. Shou. Egocentric video-language pretraining. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[194] N. Lin, T. Ohkawa, Y. Huang, M. Zhang, M. Cai, M. Li, R. Furuta, and Y. Sato. Simhand: Mining similar hands for large-scale 3d hand pose pre-training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[195] X. Lin, F. Petroni, G. Bertasius, M. Rohrbach, S.-F. Chang, and L. Torresani. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13843–13853, 2022.

[196] D. Liu, Z. Bao, J. Mi, Y. Gan, M. Ye, and J. Zhang. Cross-domain video action recognition via adaptive gradual learning. *Neurocomputing*, 556:126622, 2023.

[197] R. Liu, R. Chen, A. Abuduweili, and C. Liu. Proactive human-robot co-assembly: Leveraging human intention prediction and robust safe control. In *Proceedings of the IEEE Conference on Control Technology and Applications (CCTA)*, pages 1003–1008, 2023.

[198] R. Liu, T. Ohkawa, M. Zhang, and Y. Sato. Single-to-dual-view adaptation for egocentric 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 677–686, 2024.

[199] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 9264–9275, 2023.

[200] S. Liu, H. Jiang, J. Xu, S. Liu, and X. Wang. Semi-supervised 3D hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14687–14697, 2021.

[201] Y. Liu, J. Jiang, and J. Sun. Hand pose estimation from rgb images based on deep learning: A survey. In *IEEE International Conference on Virtual Reality (ICVR)*, pages 82–89, 2021.

[202] Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[203] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[204] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022.

[205] M. Long, H. Zhu, J. Wang, and M. I. Jorda. Unsupervised domain adaptation with residual transfer networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 136–144, 2016.

[206] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics (ToG)*, 34(6):248:1–248:16, 2015.

[207] S. Lu, D. N. Metaxas, D. Samaras, and J. Oliensis. Using multiple cues for hand tracking and model refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 443–450, 2003.

[208] Y. Lu and W. W. Mayol-Cuevas. Understanding egocentric hand-object interactions from hand pose estimation. *CoRR*, abs/2109.14657, 2021.

[209] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, and

F. Zhang et al. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019.

[210] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2507–2516, 2019.

[211] L. V. D. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9:2579–2605, 2008.

[212] I. Majil, M.-T. Yang, and S. Yang. Augmented reality based interactive cooking guide. *Sensors*, 22(21):8290, 2022.

[213] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State. Isaac Gym: High performance GPU based physics simulation for robot learning. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[214] P. Mandikal and K. Grauman. DexVIP: Learning dexterous grasping with human hand pose priors from video. In *Proceedings of the Conference on Robot Learning (CoRL)*, pages 651–661, 2021.

[215] J. Martinez, E. Kim, J. Romero, T. Bagautdinov, S. Saito, S.-I. Yu, S. Anderson, M. Zollhöfer, T.-L. Wang, S. Bai, S.-E. Wei, R. Joshi, W. Borsos, T. Simon, J. Saragih, P. Theodosis, A. Greene, A. Josyula, S. M. Maeta, A. I. Jewett, S. Venshtain, C. Heilman, Y.-T. Chen, S. Fu, M. E. A. Elshaer, T. Du, L. Wu, S.-C. Chen, K. Kang, M. Wu, Y. Emad, S. Longay, A. Brewer, H. Shah, J. Booth, T. Koska, K. Haidle, J. C.-H. Hsu, T. Dauer, P. Selednik, T. Godisart, S. Ardisson, M. Cipperly, B. Humberston, L. Farr, B. Hansen, P. Guo, D. Braun, S. Krenn, H. Wen, L. Evans, N. Fadeeva, M. Stewart, G. Schwartz, D. Gupta, G. Moon, K. Guo, Y. Dong, Y. Xu, T. Shiratori, F. A. Prada Nino, B. R. Pires, B. Peng, J. Buffalini, A. Trimble, K. A. A. McPhail, M. R. Schoeller, and Y. Sheikh. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*. IEEE, 2024.

[216] R. McKee, D. McKee, D. Alexander, and E. Paillat. NZ sign language exercises. Deaf Studies Department of Victoria Uni-

versity of Wellington, http://www.victoria.ac.nz/llc/llc_resources/nzsl.

[217] L. Melas-Kyriazi and A. K. Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12435–12445, 2021.

[218] S. Melax, L. Keselman, and S. Orsten. Dynamics based 3d skeletal hand tracking. In *Proceedings of the Graphics Interface (GI)*, pages 63–70, 2013.

[219] R. Mendonca, S. Bahl, and D. Pathak. Structured world models from human videos. In *Robotics: Science and Systems (RSS)*, 2023.

[220] H. Meng, S. Jin, W. Liu, C. Qian, M. Lin, W. Ouyang, and P. Luo. 3D interacting hand pose estimation by hand de-occlusion and removal. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 380–397, 2022.

[221] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886, 2020.

[222] A. Miech, D. Zhukov, J. B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640. IEEE, 2019.

[223] A. Miller and P. Allen. Graspit!: A versatile simulator for robotic grasping. *IEEE Robotics and Automation Magazine (RAM)*, 11:110–122, 2005.

[224] K. Miura, M. Takano, R. Hamada, I. Ide, S. Sakai, and H. Tanaka. Associating semantically structured cooking videos with their preparation steps. *IEICE Transactions on Information and Systems*, 36(2):51–62, 2005.

[225] N. Miyata, M. Kouchi, T. Kurihara, and M. Mochimaru. Modeling of human hand link structure from optical motion capture data. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2129–2135, 2004.

[226] G. Moon. Bringing inputs to shared domains for 3D interacting hands recovery in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17028–17037, 2023.

[227] G . Moon, H . Choi, and K . M . Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2307–2316, 2022.

[228] G . Moon, T . Shiratori, and S . Saito. Expressive whole-body 3d gaussian avatar. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[229] G. Moon, S. Saito, W. Xu, R. Joshi, J. Buffalini, H. Bellan, N. Rosen, J. Richardson, M. Mize, P. de Bree, T. Simon, B. Peng, S. Garg, K. McPhail, and T. Shiratori. A dataset of relighted 3D interacting hands. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)Datasets and Benchmarks Track*, 2023.

[230] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee. 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)*, pages 548–564, 2020.

[231] A. Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling, and S. Narayanan. Real-time sign language detection using human pose estimation. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, pages 237–248, 2020.

[232] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5716–5726, 2017.

[233] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt. GANerated Hands for real-time 3D hand tracking from monocular RGB. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 49–59, 2018.

[234] F. Mueller, M. Davis, F. Bernard, O. Sotnychenko, M. Verschoor, M. A. Otaduy, D. Casas, and C. Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (ToG)*, 38(4):49:1–49:13, 2019.

[235] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. Real-time hand tracking under occlusion from an egocentric

RGB-D sensor. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1163–1172, 2017.

[236] L . Müller, A . A . A . Osman, S . Tang, C . P . Huang, and M . J . Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999. Computer Vision Foundation / IEEE, 2021.

[237] L . Müller, V . Ye, G . Pavlakos, M . J . Black, and A . Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9687–9697. IEEE, 2024.

[238] J. Munro and D. Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 119–129, 2020.

[239] K. Nakamura, H. Ohashi, and M. Okada. Sensor-augmented egocentric-video captioning with dynamic modal attention. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 4220–4229, 2021.

[240] S. Narasimhaswamy, T. Nguyen, M. Huang, and M. Hoai. Whose hands are these? hand detection and hand-body association in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4879–4889, 2022.

[241] N. Neverova, C. Wolf, F. Nebout, and G. W. Taylor. Hand pose estimation through semi-supervised and weakly-supervised learning. *Computer Vision and Image Understanding*, 164:56–67, 2017.

[242] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 9912, pages 483–499, 2016.

[243] E. Ng, S. Ginosar, T. Darrell, and H. Joo. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11865–11874, 2021.

[244] E. Ng, J. Romero, T. M. Bagautdinov, S. Bai, T. Darrell, A. Kanazawa, and A. Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1001–1010, 2024.

[245] A . Q . Nichol and P . Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 8162–8171, 2021.

[246] T. Nishimura, A. Hashimoto, Y. Ushiku, H. Kameko, and S. Mori. Recipe generation from unsegmented cooking videos. *CoRR*, abs/2209.10134, 2022.

[247] T. Nishimura, K. Sakoda, A. Hashimoto, Y. Ushiku, N. Tanaka, F. Ono, H. Kameko, and S. Mori. Egocentric biochemical video-and-language dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3129–3133, 2021.

[248] M. Oberweger, G. Riegler, P. Wohlhart, and V. Lepetit. Efficiently creating 3d training data for fine hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4957–4965, 2016.

[249] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a feedback loop for hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3316–3324, 2015.

[250] T. Ohkawa. AssemblyHands Toolkit. https://github.com/facebookresearch/assemblyhands-toolkit, 2023.

[251] T. Ohkawa, R. Furuta, and Y. Sato. Efficient annotation and learning for 3D hand pose estimation: A survey. *International Journal on Computer Vision (IJCV)*, 131:3193—3206, 2023.

[252] T. Ohkawa, K. He, F. Sener, T. Hodan, L. Tran, and C. Keskin. AssemblyHands: Towards egocentric activity understanding via 3D hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, 2023.

[253] T. Ohkawa, N. Inoue, H. Kataoka, and N. Inoue. Augmented cyclic consistency regularization for unpaired image-to-image translation. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 362–369, 2020.

[254] T. Ohkawa, J. Lee, S. Saito, J. Saragih, F. Prada, Y. Xu, S. Yu, R. Furuta, Y.Sato, and T. Shiratori. Generative modeling of shape-dependent self-contact human poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[255] T. Ohkawa, Y.-J. Li, Q. Fu, R. Furuta, K. M. Kitani, and Y. Sato. Domain adaptive hand keypoint and pixel localization in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–87, 2022.

[256] T. Ohkawa, T. Yagi, A. Hashimoto, Y. Ushiku, and Y. Sato. Foreground-aware stylization and consensus pseudo-labeling for domain adaptation of first-person hand segmentation. *IEEE Access*, 9:94644–94655, 2021.

[257] T. Ohkawa, T. Yagi, T. Nishimura, R. Furuta, A. Hashimoto, Y. Ushiku, and Y. Sato. Exo2egodvc: Dense video captioning of egocentric procedural activities using web instructional videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8324–8335, 2025.

[258] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–11, 2011.

[259] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Tracking the articulated motion of two strongly interacting hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1862–1869, 2012.

[260] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P. Y. Huang, S. W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[261] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, A. Agarwal, K. Slama, H. Long, F. Xiao, L. M. Palmer, N. Paduano, J. F. Sohl-Dickstein, L. de Jimenez, C. Hilton, M. Knight, A. Micheli, A. Kreps, R. Hesse, S. Cath, A. McCauley, A. J. Hilton, M. J. Chilton, R. H. Dhariwal, A. Agarwal, A. L. Neelakantan, A. N. Steiner, A. Askell, T. Schulman, J. Wu, A. J. Hilton, A. S. Chen, B. Mann, C. Zhu, C. L. Wainwright, D. M. Almeida, E. M. E. Lebrun, J. W. Chung, L. I. M. B. J. R. Jimenez, L. P. Palmer, M. B. E. G. Hilton, M. N. Knight, P. D. Mishkin, R. H. Dhariwal, R. R. Hesse, S. K. Cath, X. S. Long, X. Jiang, and Z. M. H. Zhang. Training language models to follow instructions with

human feedback. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[262] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, 2002.

[263] A. Pardo, F. C. Heilbron, J. L. Alcázar, A. K. Thabet, and B. Ghanem. MovieCuts: a new dataset and benchmark for cut type recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–685, 2022.

[264] G. Park, T.-K. Kim, and W. Woo. 3d hand pose estimation with a single infrared camera via domain transfer learning. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 588–599, 2020.

[265] J. Park, Y. Oh, G. Moon, H. Choi, and K. M. Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1486–1495, 2022.

[266] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

[267] B. Pease and A. Pease. *The definitive book of body language: The hidden meaning behind people's gestures and expressions*. Bantam, 2008.

[268] H. Pham, Z. Dai, Q. Xie, and Q. V. Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11557–11568, 2021.

[269] B . Poole, A . Jain, J . T . Barron, and B . Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net, 2023.

[270] V. Prabhu, S. Khare, D. Kartik, and J. Hoffman. SENTRY: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8558–8567, 2021.

[271] A. Prakash, R. Tu, M. Chang, and S. Gupta. 3d hand pose estimation in

everyday egocentric images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 183–202, 2024.

[272] S. Pramanick, Y. Song, S. Nag, K. Q. Lin, H. Shah, M. Z. Shou, R. Chellappa, and P. Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5285–5297, 2023.

[273] M. Qi, E. Remelli, M. Salzmann, and P. Fua. Unsupervised domain adaptation with temporal-consistent self-training for 3d hand-object joint reconstruction. *CoRR*, abs/2012.11260, 2020.

[274] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1106–1113, 2014.

[275] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. L. Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 11219, pages 142–159, 2018.

[276] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang. DexMV: Imitation learning for dexterous manipulation from human videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 13699, pages 570–587, 2022.

[277] R.-Z. Qiu, S. Yang, X. Cheng, C. Chawla, J. Li, T. He, G. Yan, D. J. Yoon, R. Hoque, L. Paulsen, G. Yang, J. Zhang, S. Yi, G. Shi, and X. Wang. Humanoid policy $\sim$ human policy. *CoRR*, abs/2503.13441, 2025.

[278] S . Raab, I . Leibovitch, P . Li, K . Aberman, O . Sorkine-Hornung, and D . Cohen-Or. MoDi: Unconditional motion synthesis from diverse data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13873–13883. IEEE, 2023.

[279] M. Rad, M. Oberweger, and V. Lepetit. Domain transfer for 3d pose estimation from color images without manual annotations. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, volume 11365, pages 69–84, 2018.

[280] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger. Learning transferable visual models from natural language supervision. In *Proceedings of the*

International Conference on Machine Learning (ICML), pages 8748–8763, 2021.

[281] I. Radosavovic, R. P. Kosaraju, R. B. Girshick, K. He, and P. Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10425–10433, 2020.

[282] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. SAM 2: Segment anything in images and videos. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[283] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard. Recent advances in robot learning from demonstration. *Annual Review of Control, Robotics, and Autonomous Systems*, 2020.

[284] J. M. Rehg and T. Kanade. Visual tracking of high DOF articulated structures: an application to human hand tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–46, 1994.

[285] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6039–6048, 2020.

[286] P. Ren, H. Sun, Q. Qi, J. Wang, and W. Huang. SRN: stacked regression network for real-time 3D hand pose estimation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019.

[287] M. Richardson, M. Durasoff, and R. Wang. Decoding surface touch typing from hand-tracking. In *The ACM Symposium on User Interface Software and Technology (UIST)*, pages 686–696, 2020.

[288] G. Rogez, J. S. Supancic III, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4325–4333, 2015.

[289] G. Rogez, J. S. Supancic III, and D. Ramanan. Understanding everyday hands in action from RGB-D images. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision (ICCV), pages 3889–3897, 2015.

[290] G. Rogez, M. Khademi, J. S. Supancic III, J. M. M. Montiel, and D. Ramanan. 3d hand pose detection in egocentric RGB-D images. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, volume 8925, pages 356–371, 2014.

[291] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[292] J. Romero, H. Kjellström, and D. Kragic. Hands in action: real-time 3d reconstruction of hands in interaction with objects. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 458–463, 2010.

[293] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):245:1–245:17, 2017.

[294] V. Rudnev, V. Golyanik, J. Wang, H. P. Seidel, F. Mueller, M. Elgharib, and C. Theobalt. Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12365–12375, 2021.

[295] C. Ryali, Y.-T. Hu, D. Bolya, C. Wei, H. Fan, P.-Y. Huang, V. Aggarwal, A. Chowdhury, O. Poursaeed, J. Hoffman, J. Malik, Y. Li, and C. Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 29441–29454, 2023.

[296] K. Saito, Y. Ushiku, and T. Harada. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2988–2997, 2017.

[297] M . S . M . Sajjadi, O . Bachem, M . Lucic, O . Bousquet, and S . Gelly. Assessing generative models via precision and recall. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 5234–5243, 2018.

[298] N. Santavas, I. Kansizoglou, L. Bampis, E. G. Karakasis, and A. Gaster-

atos. Attention! A lightweight 2d hand pose estimation approach. *CoRR*, abs/2001.08047, 2020.

[299] N. Santavas, I. Kansizoglou, L. Bampis, E. G. Karakasis, and A. Gasteratos. Attention! A lightweight 2d hand pose estimation approach. *IEEE Sensors*, 21(10), 2021.

[300] S. I. Sayed, R. Ghoddoosian, B. Trivedi, and V. Athitsos. A new dataset and approach for timestamp supervised action segmentation using human object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3133–3142, 2023.

[301] M. Schröder, J. Maycock, and M. Botsch. Reduced marker layouts for optical motion capture of hands. In *Proceedings of the ACM SIGGRAPH Conference on Motion in Games (MIG)*, pages 7–16. ACM, 2015.

[302] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[303] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21096–21106, 2022.

[304] M. S. Shamil, D. Chatterjee, F. Sener, S. Ma, and A. Yao. On the utility of 3d hand poses for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 436–454, 2024.

[305] D. Shan, J. Geng, M. Shu, and D. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9866–9875, 2020.

[306] T. Sharp, C. Keskin, D. P. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. W. Fitzgibbon, and S. Izadi. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 3633–3642, 2015.

[307] S. Shimada, V. Golyanik, P. Pérez, and C. Theobalt. Decaf: monocular deformation capture for face and hand interactions. *ACM Transactions on Graphics (ToG)*, 42(d), 2023.

[308] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *CoRR*, abs/1804.09626, 2018.

[309] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653, 2017.

[310] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[311] B . Smith, C . Wu, H . Wen, P . Peluse, Y . Sheikh, J . K . Hodgins, and T . Shiratori. Constraining dense hand surface tracking with elasticity. *ACM Transactions on Graphics (ToG)*, 39(6):219:1–219:14, 2020.

[312] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 1857–1865, 2016.

[313] H. Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4004–4012, 2016.

[314] J . Song, C . Meng, and S . Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[315] T. Soucek, J.-B. Alayrac, A. Miech, I. Laptev, and J. Sivic. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13936–13946, 2022.

[316] A. Spurr, A. Dahiya, X. Wang, X. Zhang, and O. Hilliges. Self-supervised 3D hand pose estimation from monocular RGB via contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer*

Vision (ICCV), pages 11230–11239, 2021.

[317] A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, and J. Kautz. Weakly supervised 3D hand pose estimation via biomechanical constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 211–228, 2020.

[318] A. Spurr, P. Molchanov, U. Iqbal, J. Kautz, and O. Hilliges. Adversarial motion modelling helps semi-supervised hand pose estimation. *CoRR*, abs/2106.05954, 2021.

[319] A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 89–98, 2018.

[320] S. Sridhar, F. Mueller, M. Zollhoefer, D. Casas, A. Oulasvirta, and C. Theobalt. Real-time joint tracking of a hand manipulating an object from RGB-D input. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 294–310, 2016.

[321] S. Sridhar, A. Oulasvirta, and C. Theobalt. Interactive markerless articulated hand motion tracking using RGB and depth data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2456–2463, 2013.

[322] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019.

[323] A. Swamy, V. Leroy, P. Weinzaepfel, F. Baradel, S. Galaaoui, R. Brégier, M. Armando, J.-S. Franco, and G. Rogez. SHOWMe: Benchmarking object-agnostic hand-object 3D reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1927–1936, 2023.

[324] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 581–600, 2020.

[325] D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR), pages 3786–3793, 2014.

[326] D. Tang, T.-H. Yu, and T.-K. Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3224–3231, 2013.

[327] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou. COIN: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1207–1216, 2019.

[328] K. Tango, T. Ohkawa, R. Furuta, and Y. Sato. Background mixup data augmentation for hand and object-in-contact detection. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2022.

[329] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[330] B. Tekin, F. Bogo, and M. Pollefeys. 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)*, pages 4511–4520, 2019.

[331] G . Tevet, S . Raab, B . Gordon, Y . Shafir, D . Cohen-Or, and A . H . Bermano. Human motion diffusion model. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[332] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5026–5033, 2012.

[333] J. Tompson, M. Stein, Y. LeCun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):169:1–169:10, 2014.

[334] T.-D. Truong and K. Luu. Cross-view action recognition understanding from exocentric to egocentric perspective. *CoRR*, abs/2305.15699, 2023.

[335] T. H. E. Tse, K. I. Kim, A. Leonardis, and H. J. Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR), pages 1654–1664, 2022.

[336] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4068–4076, 2015.

[337] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2017.

[338] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

[339] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision (IJCV)*, 118(2):172–193, 2016.

[340] D. Tzionas and J. Gall. A comparison of directional distances for hand pose estimation. In *German Conference on Pattern Recognition (GCPR)*, pages 131–141, 2013.

[341] A. Urooj and A. Borji. Analysis of hand segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4710–4719, 2018.

[342] D . Uthus, G . Tanzer, and M . Georg. YouTube-ASL: A large-scale, open-domain american sign language-english parallel corpus. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[343] L. O. Vasconcelos, M. Mancini, D. Boscaini, S. R. Bulò, B. Caputo, and E. Ricci. Shape consistent 2d keypoint estimation under domain shift. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 8037–8044, 2020.

[344] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[345] R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 4566–4575, 2015.

[346] M. Šarić. Libhand: A library for hand articulation, 2011. Version 0.9.

[347] T. H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pere. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2512–2521, 2019.

[348] C. Wan, T. Probst, L. V. Gool, and A. Yao. Self-supervised 3d hand pose estimation through training by fitting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10853–10862, 2019.

[349] H. Wang, B. Li, and H. Zhao. Understanding gradual domain adaptation: Improved analysis, optimal path and beyond. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 22784–22801. PMLR, 2022.

[350] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7190–7198, 2018.

[351] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(11):2740–2755, 2019.

[352] Q. Wang, L. Zhao, L. Yuan, T. Liu, and X. Peng. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3307–3317, 2023.

[353] R. Y. Wang and J. Popovic. Real-time hand-tracking with a color glove. *ACM Transactions on Graphics (ToG)*, 28(3):63, 2009.

[354] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6827–6837, 2021.

[355] T. Wang, H. Zheng, M. Yu, Q. Tian, and H. Hu. Event-centric hierarchical

representation for dense video captioning. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 31(5):1890–1900, 2021.

[356] W. Wang, Y. Wang, S. Chen, and Q. Jin. Youmakeup: A large-scale domain-specific multimodal dataset for fine-grained semantic comprehension. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5132–5142, 2019.

[357] Y. Wang, C. Peng, and Y. Liu. Mask-pose cascaded CNN for 2d hand pose estimation from single color image. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 29(11):3258–3268, 2019.

[358] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016.

[359] Y. Wen, H. Pan, T. Ohkawa, L. Yang, J. Pan, Y. Sato, T. Komura, and W. Wang. Generative hierarchical temporal transformer for hand pose and action modeling. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2024.

[360] A. Wetzler, R. Slossberg, and R. Kimmel. Rule of thumb: Deep derotation for improved fingertip detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 33.1–33.12, 2015.

[361] H. Wu, K. Chen, H. Liu, M. Zhuge, B. Li, R. Qiao, X. Shu, B. Gan, L. Xu, B. Ren, M. Xu, W. Zhang, R. Ramachandra, C.-W. Lin, and B. Ghanem. NewsNet: a novel dataset for hierarchical temporal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10669–10680, 2023.

[362] M.-Y. Wu, P.-W. Ting, Y.-H. Tang, E. T. Chou, and L.-C. Fu. Hand pose estimation in object-interaction based on deep learning for virtual reality applications. *Journal of Visual Communication and Image Representation*, 70:102802, 04 2020.

[363] C. Wuu, N. Zheng, S. Ardisson, R. Bali, D. Belko, E. Brockmeyer, L. Evans, T. Godisart, H. Ha, A. Hypes, T. Koska, S. Krenn, S. Lombardi, X. Luo, K. McPhail, L. Millerschoen, M. Perdoch, M. Pitts, A. Richard, J. M. Saragih, J. Saragih, T. Shiratori, T. Simon, M. Stewart, A. Trimble, X. Weng, D. Whitewolf, C. Wu, S. Yu, and Y. Sheikh. Multiface: A dataset

for neural face rendering. *CoRR*, abs/2207.11243, 2022.

[364] Y. Xia, X. Zhou, E. Vouga, Q. Huang, and G. Pavlakos. Reconstructing humans with a biomechanically accurate skeleton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5355–5365, 2025.

[365] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le. Unsupervised data augmentation for consistency training. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[366] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, T. Yu, J. T. Zhou, and J. Yuan. A2J: anchor-to-joint regression network for 3D articulated pose estimation from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 793–802, 2019.

[367] C. Xu and L. Cheng. Efficient hand pose estimation from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3456–3462, 2013.

[368] H . Xu, E . G . Bazavan, A . Zanfir, W . T . Freeman, R . Sukthankar, and C . Sminchisescu. GHUM & GHUML: generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, 2020.

[369] J. Xu, Y. Huang, J. Hou, G. Chen, Y. Zhang, R. Feng, and W. Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13525–13536, 2024.

[370] S . Xu, Z . Li, Y . Wang, and L . Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14882–14894. IEEE, 2023.

[371] Y. Xu, J. Yang, H. Cao, Z. Chen, Q. Li, and K. Mao. Partial video domain adaptation with partial adversarial temporal attentive network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9312–9321, 2021.

[372] Z. Xue and K. Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *CoRR*,

abs/2306.05526, 2023.

[373] T. Yagi, M. Ohashi, Y. Huang, R. Furuta, S. Adachi, T. Mitsuyama, and Y. Sato. Finebio: A fine-grained video dataset of biological experiments with hierarchical annotation. *CoRR*, abs/2402.00293, 2024.

[374] L. Yan, B. Fan, S. Xiang, and C. Pan. CMT: cross mean teacher unsupervised domain adaptation for VHR image semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[375] A. Yang, A. Nagrani, I. Laptev, J. Sivic, and C. Schmid. VidChapters-7M: video chapters at scale. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS): Datasets and Benchmarks Track*, 2023.

[376] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid. Vid2Seq: large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10714–10726, 2023.

[377] L. Yang, S. Chen, and A. Yao. Semihand: Semi-supervised hand pose estimation with consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11364–11373, 2021.

[378] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10371–10381, 2024.

[379] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. Depth anything v2. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[380] L. Yang, J. Li, W. Xu, Y. Diao, and C. Lu. Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[381] S. Yang, S. Kwak, J. Lee, and J. Kim. Beyond Instructions: a taxonomy of information types in how-to videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 797:1–797:21, 2023.

[382] Y. Ye, A. Gupta, and S. Tulsiani. What's in your hands? 3D reconstruction

of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3885–3895, 2022.

[383] Y. Ye, P. Hebbar, A. Gupta, and S. Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19660–19671, 2023.

[384] B. Yi, V. Ye, M. Zheng, L. Müller, G. Pavlakos, Y. Ma, J. Malik, and A. Kanazawa. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[385] Y. Yin, C. Guo, M. Kaufmann, J. J. Zarate, J. Song, and O. Hilliges. Hi4D: 4D instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17016–17027, 2023.

[386] Y. You, I. Gitman, and B. Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[387] Z . Yu, S . Huang, Y . Cheng, and T . Birdal. SignAvatars: A large-scale 3d sign language holistic motion dataset and benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[388] Z. Yu, Y. Huang, R. Furuta, T. Yagi, Y. Goutsu, and Y. Sato. Fine-grained affordance annotation for egocentric hand-object interaction videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2154–2162, 2023.

[389] Z. Yu, S. Zafeiriou, and T. Birdal. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[390] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Y. Chang, K. M. Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge, J. Yuan, X. Chen, G. Wang, F. Yang, K. Akiyama, Y. Wu, Q. Wan, M. Madadi, S. Escalera, S. Li, D. Lee, I. Oikonomidis, A. A. Argyros, and T-K. Kim. 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)*, pages 2636–2645, 2018.

[391] S. Yuan, B. Stenger, and T.-K. Kim. Rgb-based 3d hand pose estimation via privileged learning with depth images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[392] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim. 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)*, pages 2605–2613, 2017.

[393] A. Zala, J. Cho, S. Kottur, X. Chen, B. Oguz, Y. Mehdad, and M. Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23056–23065, 2023.

[394] O. N. Zestas, D. N. Soumis, K. D. Kyriakou, K. Seklou, and N. D. Tselikas. A computer-vision based hand rehabilitation assessment suite. *International Journal of Electronics and Communications*, 169:154762, 2023.

[395] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12104–12113, 2022.

[396] C. Zhang, G. Wang, X. Chen, P. Xie, and T. Yamasaki. Weakly supervised segmentation guided hand pose estimation during interaction with unknown objects. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, pages 2673–2677, 2020.

[397] H. Zhang, W. Chen, Y. Wang, S. Kuo, G. Wetzstein, J. Song, and O. Hilliges. Artigrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 235–246, 2024.

[398] H. Zhang, S. Christen, Z. Fan, O. Hilliges, and J. Song. Graspxl: Generating grasping motions for diverse objects at scale. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 386–403, 2024.

[399] J. Zhang, J. Deng, C. Ma, and R. A. Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[400] L. Zhang, S. Zhou, S. Stent, and J. Shi. Fine-grained egocentric hand-

object segmentation: Dataset, model, and applications. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 127–145, 2022.

[401] M. Zhang, X. Cheng, D. Copeland, A. D. Desai, M. Y. Guan, G. A. Brat, and S. Yeung. Using computer vision to automate hand detection and tracking of surgeon movements in videos of open surgery. In *American Medical Informatics Association Annual Symposium*, 2020.

[402] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. Zheng. End-to-end hand mesh recovery from a monocular RGB image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2354–2364, 2019.

[403] Y. Zhang, L. Chen, Y. Liu, W. Zheng, and J. Yong. Adaptive wasserstein hourglass for weakly supervised RGB 3d hand pose estimation. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pages 2076–2084, 2020.

[404] Y. Zhao, T. Kwon, P. Streli, M. Pollefeys, and C. Holz. EgoPressure: A dataset for hand pressure and pose estimation in egocentric vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[405] X. Zheng, C. Wen, Z. Xue, P. Ren, and J. Wang. Hamuco: Hand pose estimation via multiview collaborative self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20763–20773, 2023.

[406] B. Zhou, H. Yuan, Y. Fu, and Z. Lu. Learning diverse bimanual dexterous manipulation skills from human demonstrations. *CoRR*, abs/2410.02477, 2024.

[407] L. Zhou, C. Xu, and J. J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7590–7598, 2018.

[408] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8739–8748, 2018.

[409] X. Zhou, A. Karpur, C. Gan, L. Luo, and Q. Huang. Unsupervised domain

adaptation for 3d keypoint estimation via view consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 11216, pages 141–157, 2018.

[410] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei. Model-based deep hand pose estimation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2421–2427, 2016.

[411] Y . Zhou, C . Barnes, J . Lu, J . Yang, and H . Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019.

[412] Y. Zhou, M. Habermann, W. Xu, I. Habibie, C. Theobalt, and F. Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5345–5354, 2020.

[413] Z. Zhou. SimpleHand: Winner of the HANDS'2023 AssemblyHands Challenge @ ICCV. https://github.com/patienceFromZhou/simpleHand, 2024.

[414] Z. Zhou, Z. Lv, S. Zhou, M. Zou, T. Wu, M. Yu, Y. Tang, and J. Liang. 1st place solution of egocentric 3D hand pose estimation challenge 2023 technical report: A concise pipeline for egocentric hand pose reconstruction. *CoRR*, abs/2310.04769, 2023.

[415] Z. Zhou, S. Zhou, Z. Lv, M. Zou, Y. Tang, and J. Liang. A simple baseline for efficient hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1367–1376, 2024.

[416] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[417] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[418] A. Ziani, Z. Fan, M. Kocabas, S. J. Christen, and O. Hilliges. Tempclr: Reconstructing hands via time-coherent contrastive learning. In *Proceedings*

of the International Conference on 3D Vision (3DV), pages 627–636, 2022.

[419] C. Zimmermann, M. Argus, and T. Brox. Contrastive representation learning for hand shape estimation. In *Proceedings of the DAGM German Conference on Pattern Recognition (GCPR)*, volume 13024, pages 250–264, 2021.

[420] C. Zimmermann and T. Brox. Learning to estimate 3D hand pose from single RGB images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4913–4921, 2017.

[421] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. J. Argus, and T. Brox. 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)*, pages 813–822, 2019.

[422] Y. Zou, Z. Yu, B. V. Kumar, and J. Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018.

Publications

Publications Related to the Dissertation

Note that * indicates co-first authors.

(1) T. Ohkawa, R. Furuta, and Y. Sato. Efficient annotation and learning for 3D hand pose estimation: A survey. *International Journal on Computer Vision (IJCV)*, 131:3193–3206, 2023.

(2) Z. Fan*, T. Ohkawa*, L. Yang*, N. Lin, Z. Zhou, S. Zhou, J. Liang, Z. Gao, X. Zhang, X. Zhang, F. Li, Z. Liu, F. Lu, K. A. Zeid, B. Leibe, J. On, S. Baek, A. Prakash, S. Gupta, K. He, Y. Sato, O. Hilliges, H. J. Chang, and A. Yao. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 428–448, 2024.

(3) T. Ohkawa, J. Lee, S. Saito, J. Saragih, F. Prada, Y. Xu, S. Yu, R. Furuta, Y. Sato, and T. Shiratori. Generative modeling of shape-dependent self-contact human poses. To appear in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

(4) N. Lin*, T. Ohkawa*, M. Zhang, Y. Huang, M Cai, M. Li, R. Furuta, and Y. Sato. SiMHand: Mining similar hands for large-scale 3D hand pose pre-training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

(5) T. Ohkawa, Y.-J. Li, Q. Fu, R. Furuta, K. M. Kitani, and Y. Sato. Domain adaptive hand keypoint and pixel localization in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–87, 2022.

(6) T. Ohkawa, T. Yagi, T. Nishimura, R. Furuta, A. Hashimoto, Y. Ushiku, and Y. Sato. Exo2EgoDVC: Dense video captioning of egocentric procedural activities using web instructional videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Other Publications

Note that † indicates highly relevant publications aligned with the scope of this dissertation.

Conference Proceedings

(7) T. Ohkawa, K. He, F. Sener, T. Hodan, L. Tran, and C. Keskin. AssemblyHands: Towards egocentric activity understanding via 3D hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, 2023. **EgoVis Distinguished Paper Award at CVPR 2025**. †

(8) R. Liu, T. Ohkawa, M. Zhang, and Y. Sato. Single-to-dual-view adaptation for egocentric 3D hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 677–686, 2024. †

(9) Y. Wen, H. Pan, T. Ohkawa, L. Yang, J. Pan, Y. Sato, T. Komura, and W. Wang. Generative hierarchical temporal transformer for hand pose and action modeling. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, Part XIV, pages 49–67, 2024.

(10) T. Ohkawa, N. Inoue, H. Kataoka, and N. Inoue. Augmented cyclic consistency regularization for unpaired image-to-image translation. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 362–369, 2020.

Journal Articles

(11) T. Ohkawa, T. Yagi, A. Hashimoto, Y. Ushiku, and Y. Sato. Foreground-aware stylization and consensus pseudo-labeling for domain adaptation of first-person hand segmentation. *IEEE Access*, 9:94644–94655, 2021. †

Preprints / Under Review Papers

(12) Z. Zhou*, T. Ohkawa*, G. Zhou, T. Hirose, K. Goto, Y. Sekikawa, and N. Inoue. DF-Mamba: Deformable state space modeling for 3D hand pose estimation in interactions. Under review, 2025. †

(13) P. Chua, C. M. Fang, T. Ohkawa, R. Kushalnagar, S. Nanayakkara, and P. Maes. EmoSign: A multimodal dataset for understanding emotions in american sign language. *CoRR*, abs/2505.17090, 2025.

(14) R. Liu, T. Ohkawa, T. H. E. Tse, M. Zhang, A. Yao, and Y. Sato. Leveraging RGB images for pre-training of event-based hand pose estimation. Under review, 2025.

Workshop Presentations / Extended Abstracts

(15) T. Ohkawa, T. Yagi, T. Nishimura, R. Furuta, A. Hashimoto, Y. Ushiku, and Y. Sato. Exo2EgoDVC: Dense video captioning of egocentric procedural activities using web instructional videos. In *JST ASPIRE International Workshop on Human-Centered Vision and Media Technologies (HCVM)*, 2025.

(16) N. Lin, T. Ohkawa, M. Zhang, Y. Huang, R. Furuta, and Y. Sato. Pre-training for 3D hand pose estimation with contrastive learning on large-scale hand images in the wild. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2024.

(17) R. Liu, T. Ohkawa, M. Zhang, and Y. Sato. Single-to-dual-view adaptation for egocentric 3D hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

(18) T. Ohkawa, T. Yagi, T. Nishimura, R. Furuta, A. Hashimoto, Y. Ushiku, and Y. Sato. Exo2EgoDVC: Dense video captioning of egocentric procedural activities using web instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

(19) T. Ohkawa. AssemblyHands benchmark and challenge for egocentric 3D hand pose estimation. In *JST ASPIRE International Workshop on Human-Centered Vision and Media Technologies (HCVM)*, 2024.

(20) T. Ohkawa, K. He, F. Sener, T. Hodan, L. Tran, and C. Keskin. AssemblyHands: Towards egocentric activity understanding via 3D hand pose estimation. In *International Computer Vision Summer School (ICVSS)*, 2023.

(21) T. Ohkawa, K. He, F. Sener, T. Hodan, L. Tran, and C. Keskin. AssemblyHands: Towards egocentric activity understanding via 3D hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

(22) T. Ohkawa, Y.-J. Li, Q. Fu, R. Furuta, K. M. Kitani, and Y. Sato. Domain adaptive hand keypoint and pixel localization in the wild. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2022.

(23) K. Tango, T. Ohkawa, R. Furuta, and Y. Sato. Background mixup data augmentation for hand and object-in-contact detection. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2022.

Domestic Conferences

(24) T. Banno, T. Ohkawa, R. Liu, R. Furuta, and Y. Sato. AssemblyHands-X: 3D hand-body co-registration for understanding bi-manual human activities. In *Meeting on Image Recognition and Understanding (MIRU)*, 2025. †

(25) N. Lin*, T. Ohkawa*, M. Zhang, Y. Huang, M Cai, M. Li, R. Furuta, and Y. Sato. SiMHand: Mining similar hands for large-scale 3D hand pose pre-training. In *Meeting on Image Recognition and Understanding (MIRU)*, 2025.

(26) Y. Maeda, T. Ohkawa, R. Furuta, and Y. Sato. Hand-active object and second object detection by extending human-object interaction transformer. In *IEICE General Conference*, 2025.

(27) R. Liu, T. Ohkawa, M. Zhang, and Y. Sato. Single-to-dual-view adaptation for egocentric 3D hand pose estimation. In *Forum on Information Technology (FIT)*, 2024.

(28) T. Ohkawa, Y.-J. Li, Q. Fu, R. Furuta, K. M. Kitani, and Y. Sato. Domain adaptive hand keypoint and pixel localization in the wild. In *Meeting on Image Recognition and Understanding (MIRU)*, 2022.

(29) K. Tango, T. Ohkawa, R. Furuta, and Y. Sato. Background mixup data augmentation for hand and object-in-contact detection. In *IEICE Technical Report (PRMU)*, 2022.

(30) A. Kobayashi*, H. Tsunashima*, T. Ohkawa, H. Aizawa, Y. Qiu, H. Kataoka, and S. Morishima. Verification of cyclical annealing for object-oriented representation learning. In *IEICE Technical Report (PRMU)*, 2021.

(31) T. Ohkawa, T. Yagi, A. Hashimoto, Y. Ushiku, and Y. Sato. Foreground-aware stylization and consensus pseudo-labeling for domain adaptation of first-person hand segmentation. In *Meeting on Image Recognition and Understanding (MIRU)*, 2021. **MIRU Student Encouragement Award**.

(32) T. Ohkawa, T. Yagi, and Y. Sato. Style adapted database: Generalizing hand segmentation via semantics-aware stylization. In *IEICE Technical Report (PRMU)*, 2020. **PRMU Best Presentation of the Month**.

(33) T. Ohkawa, N. Inoue, H. Kataoka, and N. Inoue. Consistency regularization using data augmentation for cycle-consistent GANs. In *Meeting on Image Recognition and Understanding (MIRU)*, 2020.

(34) H. Tsunashima, T. Ohkawa, H. Aizawa, H. Kataoka, and S. Morishima. Stabilizing object-aware representation learning with cyclic annealing on KL regularization. In *Meeting on Image Recognition and Understanding (MIRU)*, 2020.