Bimanual Dexterity: 3D Object Reconstruction and Cross-Embodiment Learning for Generalizable Manipulation

Zehan Ma

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2025-98
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-98.html

May 16, 2025

Copyright © 2025, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Bimanual Dexterity: 3D Object Reconstruction and Cross-Embodiment Learning for Generalizable Manipulation

by

Zehan Ma

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Ken Goldberg, Chair
Professor Angjoo Kanazawa

Spring 2025

Docusign Envelope ID: 983707AD-F75B-4524-B462-E521A12C642C

Bimanual Dexterity: Enhancing 3D Object Reconstruction and Cross-Embodiment Learning for Generalizable Manipulation

by Zehan Ma

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Ken Goldberg
Research Advisor

12 MAY 2025

(Date)

* * * * * * *

Signed by:
Angjoo Kanazawa
B2525494E7F848C...

Professor Angjoo Kanazawa
Second Reader

5/13/2025

(Date)

Bimanual Dexterity: 3D Object Reconstruction and Cross-Embodiment Learning for Generalizable Manipulation

Copyright 2025
by
Zehan Ma

Abstract

Bimanual Dexterity: 3D Object Reconstruction and Cross-Embodiment Learning for Generalizable Manipulation

by

Zehan Ma

Master of Science in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Ken Goldberg, Chair

Bimanual robots are capable of performing complex tasks that require coordination and dexterity, such as folding, handovers, and assembly. In addition to their utility in task execution, bimanual platforms also offer unique advantages for generating data to support scalable perception and policy learning. This thesis explores how dual-arm robots can be leveraged to support generalizable manipulation through two complementary systems. To address the challenge of creating complete 3D object models suitable for downstream tasks, we present a method that uses coordinated in-hand scanning and regrasping to produce high-fidelity, occlusion-free 3D Gaussian Splat reconstructions from a fixed camera. Meanwhile, to overcome the scarcity of bimanual training data, we introduce a cross-embodiment learning framework that trains dual-arm policies from single-arm teleoperation, using role alternation and vision-based synthesis to generate full bimanual demonstrations. Together, these systems demonstrate how bimanual robots can facilitate scalable data generation in both perception and policy learning, reducing reliance on specialized hardware and manual supervision while improving generalization across tasks and embodiments.

To those who stood by me quietly, steadily, and with love

Contents

List of Figures

List of Tables

Acknowledgments

I can hardly believe that my master's journey has come to an end. I'm deeply grateful to everyone who supported me along the way.

First, I want to thank Professor Ken Goldberg for welcoming me into AUTOLab in Spring 2023. His guidance has shaped not only my research but also how I think and work. At AUTOLab, I had the chance to explore a wide range of exciting projects—from agricultural robotics to manipulation, robot learning, and 3D vision. I've learned so much during this time.

Thank you to Professor Angjoo Kanazawa for serving as the second reader on my thesis, and to Professor Brian Barsky, who introduced me to research through URAP and encouraged me early on.

I'm thankful to all AUTOLab members over the past two years for creating such a supportive and collaborative environment. In particular, Lawrence Yunliang Chen, Simeon Adebola, Justin Kerr, and Chung Min Kim were incredible mentors who guided me through multiple projects, patiently answered my questions, and helped me find my footing in a new field—their mentorship played a key role in my growth. I'm especially thankful to Tianshuang Qiu and Karim El-Refai for our collaboration on Omni-Scan; I still remember the late nights we spent getting everything to work, and I learned so much from both of you. I also want to thank Eric Kaiyuan Chen, Shuangyu Xie, Max Fu, Raven Huang, Roy Tsong-Yi Lin, Andrew Goldberg, Kavish Kondap, Sandeep Bajamahal, and Harsha Polavaram for being amazing collaborators and friends—working with you made research feel less lonely and a lot more fun.

I'm grateful to my friends—Zhaozhilin Yan, Ruhao Pang, Wan Jiang, Naixiang Gao, and Yunyang Zhang—for standing by me throughout this journey. I especially want to thank Jingcong Chen for being there during the most stressful and difficult times—your companionship and support helped me through.

Finally, I owe my greatest thanks to my family—especially my parents, my sister, and my brother-in-law. Your unwavering love, belief in me, and endless support gave me the strength to persevere through every challenge. I couldn't have made it here without you.

Chapter 1

Introduction

Bimanual manipulation is essential for many real-world robotic tasks—such as folding, handovers, packing, and assembly—where dual-arm coordination offers enhanced dexterity, stability, and control. Although bimanual platforms are receiving increased attention, their adoption remains limited by challenges in both perception and learning. In perception, 3D object models are essential for applications like simulation, inspection, and policy finetuning. However, most scanning systems rely on multi-camera arrays, laser scanners, or wrist-mounted cameras, which are costly, constrained in workspace, and often suffer from occlusion. On the learning side, progress on bimanual policy training is bottlenecked by the scarcity of hardware and the difficulty of collecting high-quality demonstrations. These typically require dual-arm teleoperation or task-specific programming, making them hard to scale. These challenges point to a broader need: how can bimanual robots be used not just for execution, but also as platforms for scalable data generation in both perception and policy learning?

In Chapter 3 of this thesis, we introduce Omni-Scan, a 3D reconstruction system that uses a bimanual robot to generate complete, high-quality object models without moving cameras. The robot performs in-hand scanning with mid-air handovers, allowing each surface of the object to be captured from a fixed external viewpoint. We integrate visual segmentation and optical flow to isolate the object from the grippers and background and train high-fidelity 3D Gaussian Splat representations from the merged views. Our system achieves accurate and complete reconstructions, which we apply to the task of automated defect detection.

In Chapter 4, we present MonoDuo, a cross-embodiment learning framework that trains bimanual manipulation policies using only single-arm demonstrations. We design a teleoperation protocol where a human alternates control between the left and right sides of a bimanual task. From this complementary dataset, we synthesize full bimanual demonstrations using vision-based hand tracking, point cloud fusion, and inpainting. These demonstrations are grounded in the kinematics of the target bimanual robot and used to train policies that generalize across hardware configurations. Our evaluation shows that MonoDuo enables zero-shot deployment on unseen dual-arm robots, and improves significantly with few-shot finetuning.

In addition to the two primary systems explored in this thesis, I have worked on several other projects aimed at improving the scalability and generalization of robotic systems. In DROID and Open X-Embodiment [4, 5], I collaborated on the development of large-scale, diverse robot manipulation datasets collected across hundreds of robots, environments, and tasks. While these datasets focus on single-arm demonstrations, they provide a valuable foundation for future extensions using frameworks like MonoDuo to synthesize bimanual data, enabling cross-embodiment policy learning at scale. I also contributed to BloxNet [6], a generative design-for-robot-assembly system that transforms text prompts into physical structures assembled by a robot. While the system primarily focuses on high-level design and single-arm execution, the underlying assembly process could be extended to benefit from dual-arm coordination—particularly for handling larger components, stabilizing parts, or enabling more complex spatial arrangements. Similarly, the Gasket Assembly [7], which investigates long-horizon, high-precision gasket insertion tasks, reflects the challenges of contact-rich manipulation where bimanual strategies—such as alignment with one arm and insertion with the other—could further improve reliability and success.

Together, these contributions demonstrate how bimanual robots can serve not only as capable manipulation agents, but also as platforms for scalable data generation—supporting both perception and policy learning. While this thesis primarily focuses on addressing two key bottlenecks—occlusion-free 3D object modeling and data-efficient bimanual policy training—it is situated within a broader research context that explores generalization, coordination, and scalable learning in robotics. These efforts collectively point toward more autonomous and generative robotic systems.

Chapter 2

Related Work

2.1 3D Scanning and Part Inspection

3D Reconstruction with Radiance Fields

Neural Radiance Fields [8] are an attractive representation for high quality scene reconstruction from posed RGB images, with a flurry of recent work enhancing quality [9, 10, 11, 12], large-scale scenes [13, 14, 15], optimization speed [16, 17, 18, 19], dynamic scenes [20, 21, 22], and more. Because of its high-quality reconstruction and differentiable properties, NeRF has been explored in robotics for navigation and mapping [9, 23, 24, 25], manipulation [26, 27, 28, 29, 30, 31, 32], and for synthetic data generation [33]. 3D Gaussian Splatting [34] made a major breakthrough in speed and quality of radiance fields, and the field has quickly adopted it for similar applications. In this work we use 3DGS to reconstruct high-quality object models, and in contrast to prior work reconstruct entire objects in high detail with a static camera, via a method of merging multiple scans and accurately masking the object of interest.

3D Object Scan Data

Conventionally, datasets of 3D objects are constructed with expensive equipment like multiview camera arrays or high precision depth sensors, such as in the Google Scanned Objects [35] or DTU [36] datasets. Other large datasets like Objaverse [37] exist, but are comprised of synthetic objects. In this work, we leverage recent work on multi-view reconstruction from RGB images to alleviate the need for expensive sensors and autonomously digitize real objects with a robot.

Several works explore reconstructing objects in human hands, including Color-NeuS [38], which reconstructs object SDFs by separating view-dependent effects with a relighting network. BundleSDF [39] achieves near real-time tracking and reconstruction from monocular RGB-D video through pose graph optimization.

Automated Part Inspection

Automated part inspection using robotics has advanced significantly with vision systems, machine learning, and sensor integration. Prior work has studied photogrammetry-based 3D reconstruction for inspection where the robot moves a camera around the object on a tabletop [40]. Davtalab et al. (2022) developed a deep learning approach for real-time defect detection in additive manufacturing, improving quality control [41]. In this work, we focus on small parts like plug adaptors or cameras, and leverage a dual handover grasp to automatically inspect small parts holistically without occlusions from a tabletop or gripper.

2.2 Cross-Embodiment and Bimanual Manipulation

Learning-Based Approaches to Bimanual Manipulation

Existing learning-based approaches to equip robots with bimanual manipulation skills can be broadly classified into three categories: learning from demonstrations [42, 43, 44, 45, 46, 47, 48, 49], sim-to-real reinforcement learning [50, 51, 52], and learning from human videos [53, 54, 55, 56, 57, 58, 59] or human motion data [60, 61]. In learning from demonstrations, a human controls the robot arms and end-effectors through kinesthetic teaching or a teleoperation system of choice, so that sensorimotor data can be directly collected from the bimanual robot. The collected data can then be used to train bimanual manipulation policies in a straightforward way, using state-of-the-art imitation learning policies [62]. Advances in bimanual teleoperation system [44, 46, 47, 49] and imitation learning algorithms [63, 64] in recent years have lowered the barrier for adopting this approach, making it a popular choice among both industry and academic labs. Sim-to-real reinforcement learning approaches, on the other hand, typically do not use any real-world human or robot data. Existing works [50, 51, 52] instead first learns bimanual policies using the "digital twin" of the target robot in a simulator of choice, then transfer the learned policy to the real bimanual robot. This often leads to challenges in reward design and sim-to-real gap. Learning from human video or motion data falls between the previous two approaches, in that it eases bimanual policy learning by learning directly from human action priors but does not contain robot action data that can be directly used. Our work is most closely related to learning from human video, which we discuss below.

Cross-Embodiment Robot Learning

Cross-embodiment robot learning [65] aims to learn or transfer policies across robots with different physical embodiments. This is crucial for generalizing learning across hardware platforms, reducing the need to retrain for every new robot configuration. A growing line of research tackles this problem using domain randomization to learn robot-conditioned policies [66, 67, 68, 69, 70, 71, 72, 73, 74, 75] or training on large real robot data [76, 77, 78, 79, 80, 81, 82, 83] to make policies more robust and generalizable [84, 85, 86, 87, 88, 89,

90, 91, 92, 93, 94, 95, 96, 97, 98]. RoVi-Aug [99] introduces a diffusion-based augmentation pipeline that replaces the robot in demonstration images and generates new camera views, producing synthetic demonstrations with varied robots and viewpoint. In contrast to offline augmentation, Mirage [100] performs test-time image editing to create an illusion that the original robot is performing the task. This "cross-painting" technique decouples visual differences from control and achieved successful zero-shot transfer of manipulation policies between different single-arm embodiments. We build on these cross-painting techniques in this work. A related approach, Shadow [101], simplifies cross-embodiment image editing by overlaying segmentation masks of the source and target robots on training and test images.

Learning from human video can be seen as a special form of cross-embodiment robot learning. Several recent works that visually translate across embodiments leverage human demonstration videos as a source of robot training data. For example, Phantom [102] allows training robots without any robot demonstrations, by converting human videos into robot-like observation; EgoMimic [103] co-trains policies on egocentric human videos and matched teleoperated robot demonstrations with cross-domain alignment. These methods show that visual editing and alignment can make human video demonstrations viable for robot policy learning. Other existing frameworks [104, 84, 105, 55] also explore translating human videos into robot actions via learned correspondences.

Learning Bimanual Manipulation with A Single-Arm Robot

Bimanual manipulation presents additional challenges for cross-embodiment learning due to coordinated, high-dimensional actions. Prior multi-embodiment studies usually avoid direct visual retargeting or handle bimanual robots with specialized architectures. For example, CrossFormer [106] demonstrates a single policy controlling a bimanual robot across embodiments with a separate action head for bimanual robots. DexMimicGen [107] tackles bimanual learning by leveraging a small set of teleoperated demonstrations to seed large-scale synthetic trajectory generation in simulation, where the demonstrations need to adhere to dual-arm execution trajectories on the same robot. AnyBimanual [108] addresses bimanual learning by composing single-arm skills through a task-level reasoning and planning module, guided by a few task-specific bimanual demonstrations.

In contrast, our approach learns end-to-end coordinated bimanual policies directly from synthetic demonstrations generated using only single-arm robot data paired with human interaction. The only closely related work that we have found is LfDT [109], which uses human-robot interaction videos to learn dual-arm policies by learning a CycleGAN to transform human-robot images into robot-robot images. However, LfDT requires robot-robot target domain videos for training the CycleGAN [110] and is only validated on relatively simple tasks such as pushing. To the best of our knowledge, this work is the first to learn bimanual manipulation policies using only a single-arm robot, and to demonstrate success on complex, contact-rich tasks with zero-shot success, as shown in Section 4.4.

Chapter 3

Omni-Scan: Bi-Manual 3D Object Reconstruction

3.1 Introduction

"Digital Twins"—visually accurate 3D reconstructions of physical objects—are useful for applications such as automated inspection in manufacturing and Sim2Real learning. Recent advances in 3D reconstruction, such as Neural Radiance Fields (NeRF) [8] and 3D Gaussian Splatting (3DGS) [34], have enabled high-quality novel view synthesis and 3D reconstruction from 2D images. Yet in robotic contexts, prior work typically relies on moving wrist-mounted cameras, which limit coverage due to kinematic constraints and inability to observe object regions near the support surface.

In this work, we present Omni-Scan, a fully autonomous system for 3D object reconstruction uses a bimanual robot to perform in-hand scanning and regrasping, requiring only a stationary RGB camera and a stereo depth sensor. The system grasps objects and rotates them in front of the camera to capture multi-view images. It then performs a bi-manual handover to re-grasp the object, exposing surfaces previously occluded by the gripper. This enables the construction of an omni-directional (360°) 3DGS model of the object suitable for downstream applications such as automated part inspection.

While a single-arm robot could theoretically place the object down on a surface and attempt a second grasp, this approach suffers from multiple drawbacks. First, regrasping on a support surface would require the second grasp to avoid all regions contacted during the first grasp to ensure occlusion-free (360°) object models. Achieving this requires the grasp planner to generate contact configurations that are both kinematically feasible and mutually disjoint. These requirements increase the complexity of the grasp planning pipeline and reduce the robustness of the overall scanning system. Second, 3D pointcloud registration has been a long unsolved question. Algorithms such as ICP require a good initial alignment to converge to the correct registration result [111], which is not available when the object is regrasped from arbitrary poses on a surface. Global registration algorithms such as [112] and

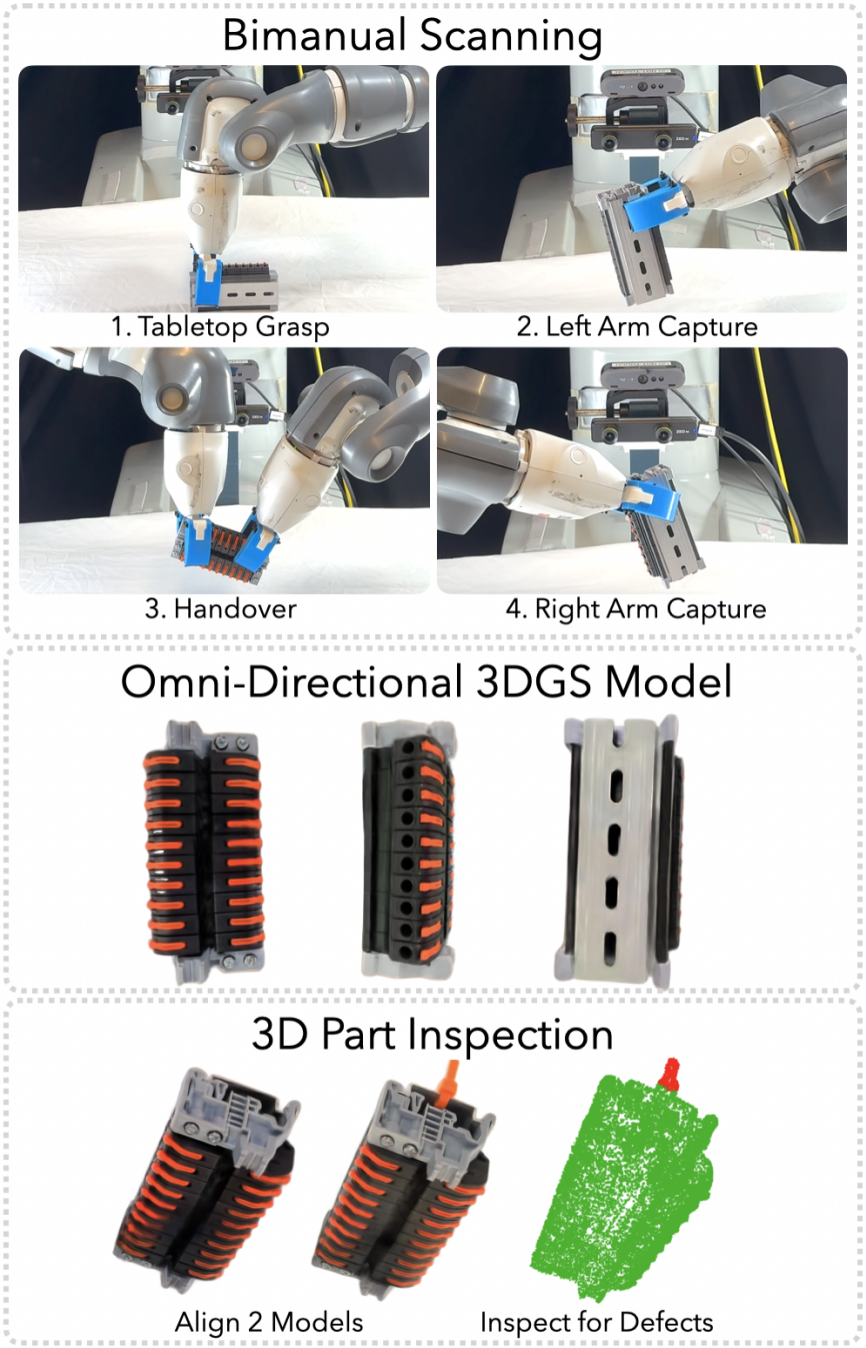

Figure 3.1: The robot grasps an object (wire connector) in any position and orientation for inspection. OMNI-SCAN then transfers the object between grippers to create a complete scan. The resulting full surface 3DGS model can be compared with a reference model for object inspection.

[113] aim to align point clouds under large pose differences and outliers. However, for in-hand scanning, which requires two complementary grasps to capture the complete object model, these methods often fail. Since each scan captures only a partial and occlusion-dependent view of the object, the resulting overlap between scans is often minimal and inconsistent, making it challenging for global registration to establish reliable correspondences and produce accurate alignments.

In contrast, the bimanual setup enables controlled mid-air handovers, where the relative pose between grasps is approximately known and provides a strong initialization for alignment, which can then be refined using global registration algorithms. This significantly reduces ambiguity when merging partial 3DGS reconstructions and enables efficient, occlusion-free data acquisition in a single autonomous scanning loop.

Additionally, in-gripper scanning presents unique challenges: object occlusions from the end effector, merging multiple independent scans, and the inversion of the typical neural reconstruction setup (static object, moving camera). To address this, we develop a masking pipeline using optical flow, DepthAnything V2 [114], Segment Anything [115], and SAM2 [116] to segment the object from background and gripper. We further modify the standard 3DGS training pipeline to support this scanning paradigm and apply Omni-Scan to industrial part inspection, identifying both visual and geometric defects in real-world objects.

This chapter introduces the following key contributions:

1. **Omni-Scan pipeline:** A bi-manual robotic scanning system including object grasping, multi-view scanning, handover for full surface coverage, and 3D model reconstruction.

2. **Robust masking pipeline:** A segmentation approach to accurately isolate the object from the gripper and background in RGB images.

3. **3DGS merging:** A pose optimization and model fusion method to combine multiple partial 3DGS reconstructions into a unified object model.

4. **Alignment for inspection:** A technique to align reconstructed models for identifying geometric and visual defects.

5. **Experimental validation:** Demonstration of 83% accuracy in defect detection across a set of 12 industrial and household objects.

3.2 Problem Statement

The goal is to create a visually-accurate omni-directional 3D model of a provided object, and then use this reconstruction to inspect for defects. We assume objects are rigid and cannot fit inside a 3cm diameter sphere but can fit inside a 10cm one, as well as the availability of

Figure 3.2: **Reconstructed 3D Gaussian Splats of the 3DGS-Merged Model** We show **rendered views** from reconstructed splat models of objects collected by Omni-Scan. Each object is fully reconstructed without occlusion, even though the data was collected while grasped. In addition, the models capture fine geometric and visual details such as text or notches. See our website for interactive videos of full 3D surfaces.

a bi-manual robot with parallel jaw grippers, one fixed high-resolution monocular camera, and one stereo camera. During reconstruction, a target object is placed within the reachable workspace of the robot on a tabletop. We assume the robot is able to grasp and lift the object (i.e it is not too heavy). During defect inspection, a robot is provided with 3DGS models of two reference objects and one new 3DGS model to evaluate. The system analyzes these 3 models to determine if the new model contains a defect and if so where. Defects can be geometric defects, meaning a structural deformation or flaw greater than 4.5mm in size, or visual defects, such as a scratch or a blemish greater than 2mm in size.

3.3 Omni-Scan

Omni-Scan first grasps the object from the tabletop, then while holding it mid-air, scans it by turning the object in front of a fixed camera to capture multiple viewpoints. We then perform a handover, passing the object from one gripper to another to scan it again from a new pose. After collecting the images, we process them with a combination of robot kinematics, Depth Anything, optical flow, and SAM to generate training poses and masks for 3DGS reconstruction. We then train 2 individual Gaussian Splat models (left and right) and merge them into a single, high-quality 3DGS model. We use the resulting model for part inspection by detecting defects compared to other examples of the same object.

Scanning Procedure

Tabletop Grasping

We use an ABB YuMi bi-manual robot with soft 3D-printed grippers [117] for compliant caging grasps. Objects are placed randomly within the workspace and captured by a ZED Mini RGB-D camera mounted at the robot base.

We generate a depth image of the object from the stereo image pairs using RAFT-Stereo [118] and one RGB image to generate object masks with SAM [115], filtering the masks by the known location of the table to isolate the object. The depth image is deprojected to create point clouds of both the scene and the isolated object, using DBSCAN [119] to remove noise. Contact-GraspNet [120] then generates candidate grasps on only the object point cloud, and the highest-scoring grasp is planned and executed with the left side gripper using the Jacobi motion planning software [121]. If the grasp is kinematically infeasible or would lead to a collision, the next highest scored grasp is chosen.

Scan Trajectory

After the object is grasped and lifted, the robot performs scanning by rotating the wrist of the gripper 360° in 20 evenly spaced longitudinal positions about its local z-axis. We evenly sample 5 latitudes from the z-axis between -10 and 70 degrees (equaling 100 images). Beyond these limits, occlusions from the gripper prevent the camera from clearly viewing the object. At each latitude, we collect the pose of the arm that is holding the object T and capture the corresponding 4K image I. The scanning process for one arm takes 6 minutes for 100 images.

Bi-Manual Re-Grasping

Since a portion of the object has been occluded throughout the entire first scan by the robot gripper, the robot then regrasps the object at a different position and scan one more time to capture these regions. To do this the robot moves the object to a predefined end-effector position easily reachable by the other arm. Following a very similar approach as 3.3, Omni-Scan generates grasps on the object point cloud after segmenting the robot arm by deleting depth points overlapping with the URDF model. We then choose the highest scored grasp, accounting for kinematic constraints and collisions. To regrasp the object, the right gripper encloses the object, then the left gripper is released. This right arm then repeats the same scanning process as detailed in 3.3.

Dataset Processing

Pose Processing

We first compute the camera-to-object transform for the left and right scans (100 images per scan). From our calibrated camera, we can get the transform from camera to world T_c. Since

we do not directly have the pose of an object center relative to the robot, we approximate it with the transform from the robot to the gripper. The reconstruction of the object is performed in the frame of the gripper.

For each image i, the pose from the camera to the object T_{ic} can be computed by its corresponding $T_i^{-1}T_c$, where T_i is the transform from the robot gripper (that is holding the object) to world, creating capture_L and capture_R, consisting of image-transform pairs.

Mask Processing

Our masking pipeline (Figure 3.3) robustly segments the object by systematically filtering out background elements, robot gripper, and robot arm. The pipeline consists of the following key components:

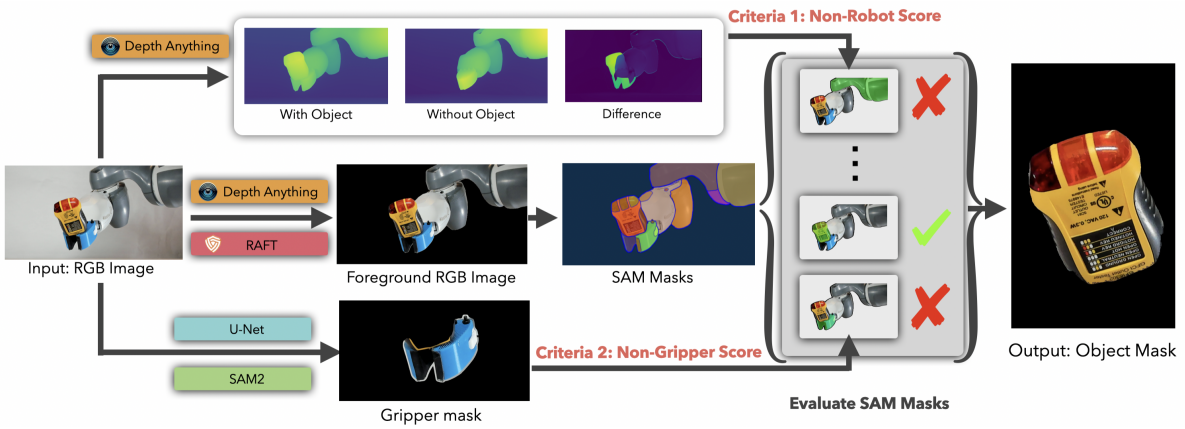

Figure 3.3: **Masking Pipeline** (1) starts with an RGB image of the robot gripper holding an object, (2) extracts the foreground to isolate potential objects, (3) uses SAM to generate candidate object masks, (4) evaluates masks using two criteria: Non-Robot Score (comparing depth with/without object) and Non-Gripper Score (using U-Net and SAM2-generated gripper masks), and (5) outputs a clean object mask containing only the target object, rejecting gripper and robot parts.

Robot Gripper Segmentation We first segment the robotic gripper to distinguish it from the object being grasped. To achieve this, we train a U-Net segmentation model using 3,000 manually labeled images for 3 objects. The ground truth labels for training U-Net were generated using SAM2 video propagation [116], where manually annotated gripper masks were propagated across frames on a training set of 3 objects. We then run inference on the scans using our trained U-Net models to obtain gripper masks. To refine U-Net masks, we select a pre-defined list of frames where the gripper is unoccluded to prompt SAM2 to generate gripper masks using video propagation.

Foreground Mask Generation To filter out background elements and distinguish the object from the robot arm, we apply:

1. **Ground Truth Dataset Depth Estimation:** We collected a ground truth dataset where no object is held inside the gripper. We then use DepthAnything V2 [114], a deep learning model for monocular depth estimation, to generate per-pixel depth predictions for each frame. The resulting depth maps are thresholded to segment foreground objects and obtain depth masks. Additionally, we save the predicted per-pixel depth values for all frames as **ground truth depth output**.

2. **Current Dataset Depth Estimation:** We use DepthAnything V2 again to get the predicted per-pixel depth values for all frames as **current depth output** for the current dataset, which will be used to compare with ground truth depth output later in our pipeline.

3. **Current Dataset Optical Flow Refinement:** We notice that DepthAnything may mistakenly classify the floor of the workspace as being close to the camera. To address this, we estimate inter-frame motion using RAFT optical flow [122]. The RAFT model computes dense optical flow by iteratively refining motion estimates at multiple scales using a correlation-based cost volume. We take the intersection of flow masks and depth masks to ensure accurate segmentation, called **foreground masks**. The combination of these masks produce **foreground-filtered images**, which will be fed into SAM in the next step.

Object Mask Generation The foreground mask is passed to SAM2 to generate a set of candidate masks. We also perform a one-time calibration where we estimate mono-depth for images from the capture trajectory without an object grasped (empty gripper). The usage of these depth maps is described next.

For each candidate mask M from SAM, we then label it as part of the robot or object based on two scoring functions:

Non-Robot Score:

$$S_{\text{NR}} = \frac{1}{|M|} \sum_{p \in M} |D_{curr}(p) - D_{empty}(p)| \tag{3.1}$$

where D_{curr} is the current depth output, D_{empty} is the empty-gripper depth output, and p represents pixels in the candidate mask. Since the depth of the robot arm in current is typically similar to the empty-gripper depth, the score helps filter out regions corresponding to the arm. In contrast, the object and gripper configuration will differ significantly from the empty gripper depth (where the gripper is fully closed and no objects are in it), resulting in a higher S_{NR} score, indicating a higher likelihood of belonging to the object.

Non-Gripper Score

$$S_{\text{NG}} = 1 - \frac{|M \cap G|}{|M|} \tag{3.2}$$

where G is the gripper mask. A higher S_{NG} score indicates less overlap with the gripper, meaning it's more likely part of the object.

We keep candidate masks with $S_{NR} \geq 150$ and $S_{NG} \geq 0.9$ as our final object mask (threshold empirically determined).

3DGS Training

After obtaining the object masks, OMNI-SCAN seeks to create one omni-directional Gaussian Splat model of the entire object without occlusions. We do this in the following steps(Figure 3.4):

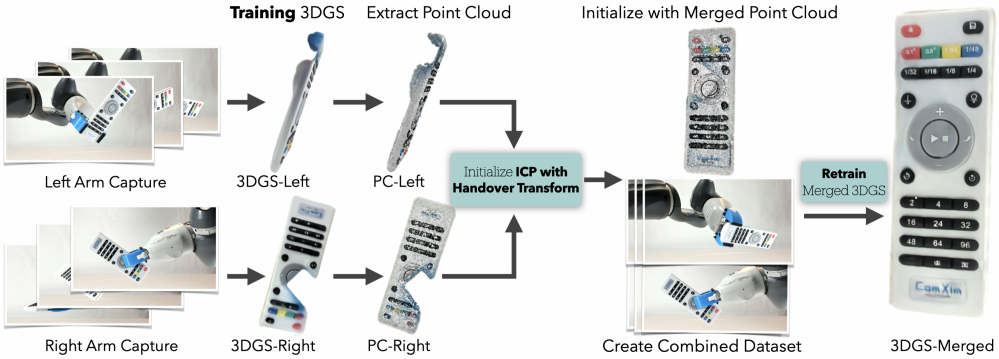

Figure 3.4: **Overview of Training Pipeline** We first train separate 3DGS models for left and right arm captures and extract their Gaussian centers as point clouds. Using the estimated handover transform T_{lr}, we initialize Iterative Closest Point (ICP) algorithm, which iteratively refines the alignment between two point clouds by minimizing the distance between corresponding points, for alignment. The refined transformation from ICP is then used to merge the datasets, enabling training of a unified 3DGS model on the combined dataset.

1. Create capture$_L$ and capture$_R$ from the left and right arm scans
2. Train Gaussian Splat models, 3DGS$_L$ and 3DGS$_R$, **individually** on capture$_L$ and capture$_R$
3. Compute capture$_{\text{merge}}$ by computing equivalent transforms between capture$_L$ and capture$_R$
4. Train 3DGS$_{\text{merged}}$ on capture$_{\text{merge}}$ as the **merged** 3D model

Compute capture$_L$ and capture$_R$

Using the method outlined in section 3.3, we compute image-transform pairs for al individual scans. This transform is still in the respective grasp frame, so it is only suitable for training the individual models, 3DGS$_L$ and 3DGS$_R$.

Training Individual Models

We first produce a 3DGS of each of the datasets individually for 16000 steps to get an estimate of the object's geometry. From these splat models, we retrieve colored point clouds P_1, P_2.

Aligning the scans to create capture$_{\text{merge}}$

To create a frame for both captures, we make use of Iterative Closest Point (ICP), an algorithm that iteratively refines the alignment between two point clouds by minimizing the distance between corresponding points. We initialize the relative point cloud transform using the transform between the two robot grippers. Let the left and right gripper positions at handover be T_{lh}, T_{rh}.

Specifically, we make the following definitions. An image taken with the left gripper is i^l, and the right gripper i^r. Its corresponding pose in its *own* frame is T^l_{ic} or T^r_{ic}.

We assign our left capture to be the canonical frame and seek to transform the right capture to the left's frame. It is necessary to compute this transformation in camera frame because 3DGS$_L$ and 3DGS$_R$ are trained with camera to gripper poses. This handover transformation in camera frame is given by $T_c^{-1}T_{lh}$ and $T_c^{-1}T_{lh}$. For each image i taken while the object is held by the right gripper, we can compute its left equivalent transform as

$$T^l_{ic} = (T_c^{-1}T_{lh})^{-1}T_c^{-1}T_{rh}T^r_{ic} \tag{3.3}$$

$$= T_{lh}^{-1}T_cT_c^{-1}T_{rh}T^r_{ic} = T_{lh}^{-1}T_{rh}T^r_{ic} \tag{3.4}$$

We use $T_{lh}^{-1}T_{rh}$ as an initialization for ICP algorithm to align the two colored point clouds P_1, P_2 extracted in 3.3. Let the optimized transform be T^*_{lr}, then the transform for images from the right scan becomes $T^l_{ic} = T^*_{lr}T^r_{ic}$. Transforms for images from the left scan remains unchanged since it is the canonical frame.

Training Omni-Directional Model on Merged Captures

Using the merged colored point clouds $P_1 + T^*_{lr}P_2$ as initialization for the 3DGS model, we train 3DGS$_{\text{merge}}$ on capture$_{\text{merge}}$ for 50000 steps.

Supporting In-Gripper Datasets

For 3DGS training we extend Nerfstudio's Splatfacto model [123, 124] to support multi-dataset training. Naively training a 3DGS on the raw image datasets is infeasible as 3DGS assumes a static scene, while our data seen from the perspective of the camera is inherently inconsistent except for the object. Thus, we must alter the losses to account for this. In addition, we must support training on datasets where the object is occluded by the gripper.

Object Opacity Loss During the training process, Gaussian Splat models produce the **accumulation** metric as well as RGB renders. The accumulation metric measures how much each pixel is covered or influenced by overlapping Gaussians during the rendering process. Accumulation quantifies the total accumulated alpha (opacity) at each pixel due to the contribution of multiple Gaussians. Lower accumulation values suggest sparse coverage, where fewer Gaussians contribute to the final pixel color. We introduce an L1 loss between the model's **accumulation** and the image's object mask, which attempts to match the rendered opacity to the calculated mask. Intuitively this penalizes any Gaussians outside of the object mask to ensure the resulting model is floater-free and has clean boundaries.

Gripper-Agnostic Losses When combining the datasets, we formulate the loss such that the model is ambivalent towards the area that the gripper occupies. Specifically, any per-pixel loss value that intersects with a gripper mask is set to 0. Importantly, this includes the previously described opacity loss, which ensures the model is able to add Gaussians that are occluded by the gripper in one dataset by analyzing the object from the other dataset's perspective.

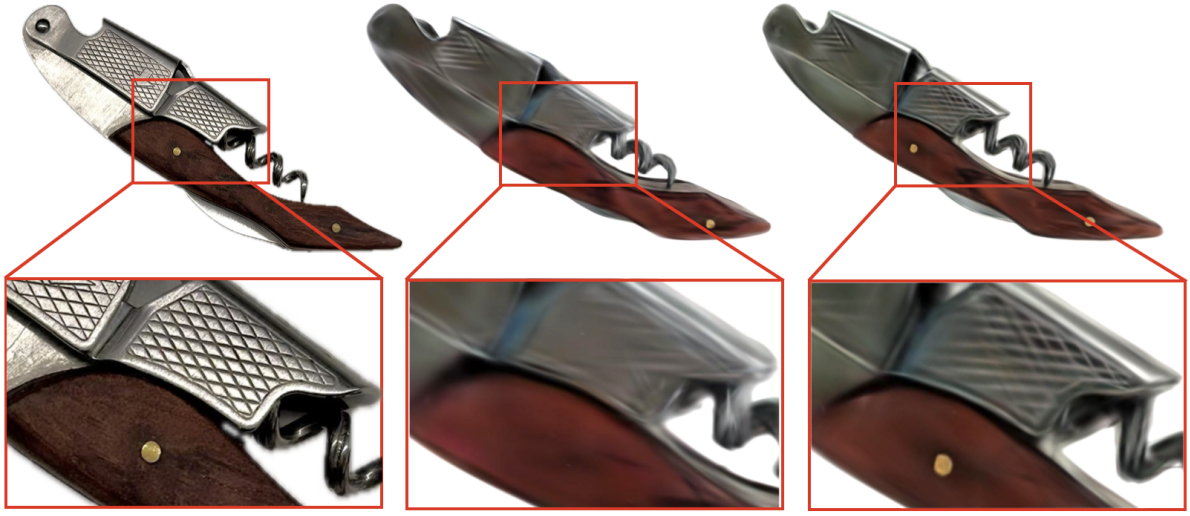

Figure 3.5: **Gripper Agnostic Loss Ablation** We perform an ablation on the Gripper Agnostic Loss, and we observe that reconstruction quality decreases without it. Specifically, the bronze stud and the cross-hatch pattern appear only when we have the Gripper Agnostic Loss.

	Realsense Camera			Remote Control			Outlet Tester			Wine Opener			Wire Connector		
	No Alignment	Handover Only	Omni-Scan	No Alignment	Handover Only	Omni-Scan	No Alignment	Handover Only	Omni-Scan	No Alignment	Handover Only	Omni-Scan	No Alignment	Handover Only	Omni-Scan
PSNR ↑	26.52	27.36	**31.12**	23.66	24.52	**26.08**	25.94	24.45	**29.26**	23.02	23.95	**30.52**	22.10	22.20	**28.51**
SSIM ↑	0.991	0.991	**0.994**	0.982	0.983	**0.984**	0.986	0.986	**0.989**	0.985	0.984	**0.989**	0.969	0.970	**0.981**
LPIPS ↓	0.015	0.015	**0.010**	0.037	0.032	**0.025**	0.019	0.018	**0.011**	0.020	0.020	**0.011**	0.046	0.038	**0.020**

Table 3.1: **Omnidirectional Object Reconstruction Quality** Comparison of reconstruction quality of 5 home, industrial, and office objects. We report metrics for each object's $3DGS_{merge}$ averaged over **200** images from the left gripper and right gripper scans. Peak Signal-to-Noise Ratio (PSNR) quantifies the quality of a reconstructed or compressed image/video by comparing it to the original on the logarithmic decibel scale, where higher values indicate better fidelity. Structural Similarity Index (SSIM) measures the similarity between two images by considering luminance, contrast, and structure, with values ranging from -1 to 1, where 1 indicates identical images. Learned Perceptual Image Patch Similarity (LPIPS) measures the perceptual similarity between images by comparing feature embeddings from a pre-trained neural network, ranging from 0 to 1, where lower values indicate higher similarity. Results suggest that Omni-Scan is able to reconstruct objects with high quality by incorporating information from all view directions.

3.4 Experiments

Physical experiments aim to evaluate 1) the quality of the 3D reconstruction, and 2) the effectiveness of the inspection system for finding defects.

Reconstruction

We collect 17 objects for reconstruction, which comprise a range of industrial, office, and household objects. We evaluate the reconstruction quality by comparing object renderings to the 200 ground truth camera images, reporting image similarity metrics (PSNR, SSIM, and LPIPS) on image regions masked by the intersection of the object mask and the accumulation (excluding the gripper) in Fig. 3.1. This penalizes accumulation and shape disparities. $3DGS_{merge}$ is compared to images from the left *and* right hand scans, ensuring that it holistically represents the object.

Results See Figure 3.2 for qualitative multi-view renders of objects reconstructed autonomously by Omni-Scan. Table 3.1 reports image quality metrics across both left and right datasets. Omni-Scan achieves high reconstruction quality, indicating it is able to reconstruct even occluded regions of the object by incorporating information from the unoccluded dataset.

Defect Inspection

We apply Omni-Scan for defect inspection on 12 distinct objects, with 3 scans for each object where 2 are of pristine reference objects and 1 contains a visual or geometric defect.

Geometric Defects	Visual Defects	Success Rate
6/7	4/5	83.3%

Table 3.2: **Correct Identifications** of the defective object using aligned pairwise comparisons.

Visual defects are changes made to the visual appearance of the object without significantly affecting its geometry. For the PVC pipe connector in Figure 3.6 we add yellow tape and mark one end of the pipe.

Geometric defects are introduced by damaging or otherwise changing the surface geometry of the object. For example, in Figure 3.6 we attach a strap to the end of the flashlight but changes such as bending, breaking, or cutting the object also qualify.

We evaluate the system's ability to identify the defective part of these 3 scans. Omni-Scan highlights the point clouds of physical defects and highlights renders of a difference visual defects. We identify the defective part using a combination of pixel-space analysis and point cloud analysis. We use TEASER++ [112], [125], a fast and robust global registration method, to obtain an initial alignment transformation between the extracted point clouds. This transformation serves as an initialization for ICP, which further refines the alignment between the Gaussian models.

Pixel Differencing We render 100 images from poses that align with the training dataset for the first dataset. Then using the alignment transform of the following 2 datasets, we compute renders of the same location and orientation. We can then directly compute the per-pixel difference of these two renders to evaluate the difference of the models. Since the two non-defective objects should be indistinguishable, we can compare pair-wise distances, and the smallest distance pair are the non-defective parts with the remainder being the defective one as demonstrated in Fig. 3.6.

Pixel Differencing Results Omni-Scan successfully detects visual defects in 4 out of 5 trials. We successfully identified defects such as scratches and tape on the pipe connector as illustrated in Fig. 3.6. Results suggest that our alignment pipeline can achieve pixel-level accuracy. The source of the failure cases is in the masking pipeline, where a portion of the gripper remains inside the object mask. This leads to artifacts in the merged 3DGS, resulting in one non-defective object being significantly different than the other non-defective one. We illustrate this in Fig. 3.8.

Point Cloud Differencing Given the aligned point clouds for any two objects, we compute the difference between them. This is done by computing the minimum distance from a point in one point cloud to any point in the other point cloud. If a point's minimum distance

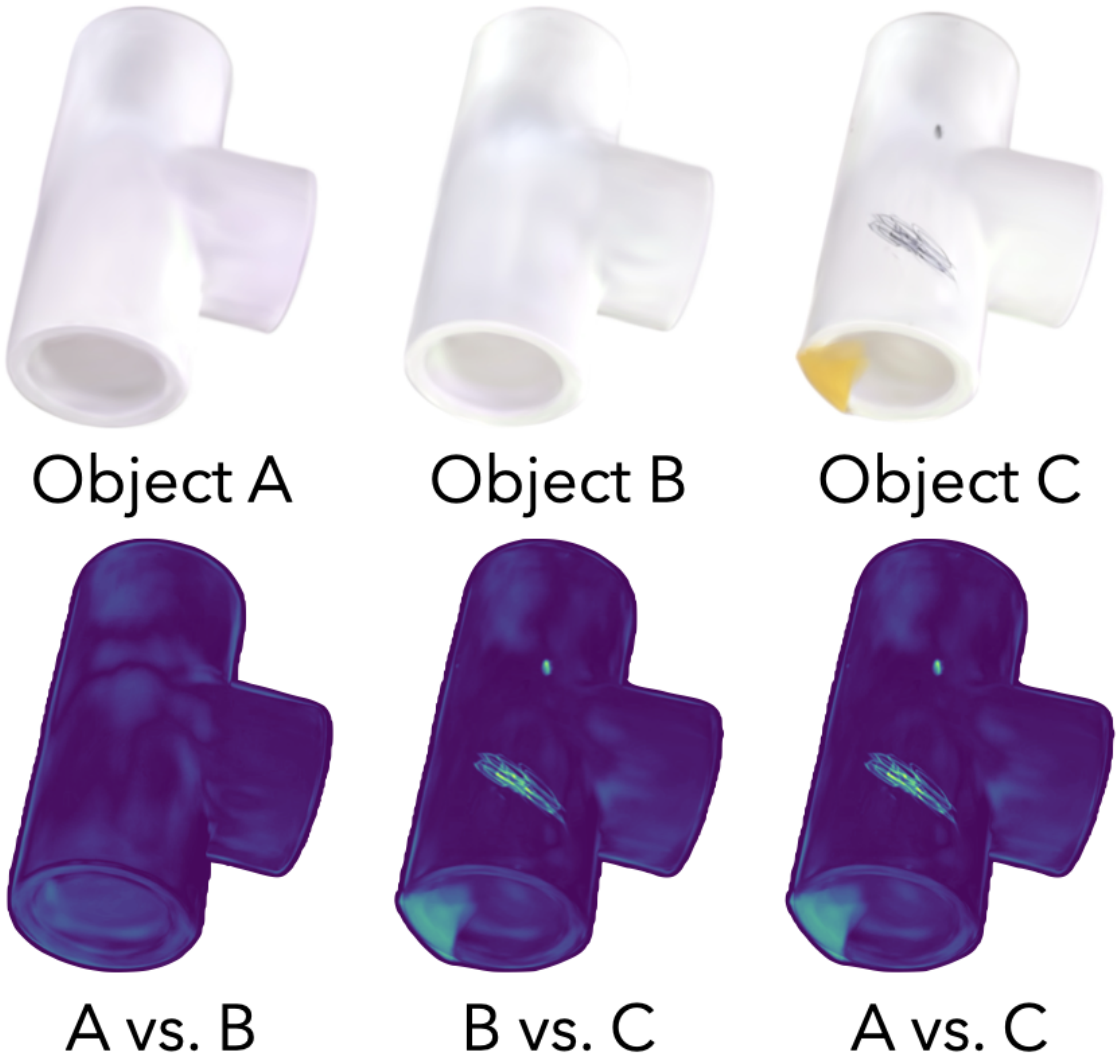

Figure 3.6: **Visual Defect Detection** *Top Row:* The rendered RGB of three Omni-Scan models. *Bottom Row:* The colorized per-pixel difference after alignment. The highest difference appears in the exact position of the scratch and tape.

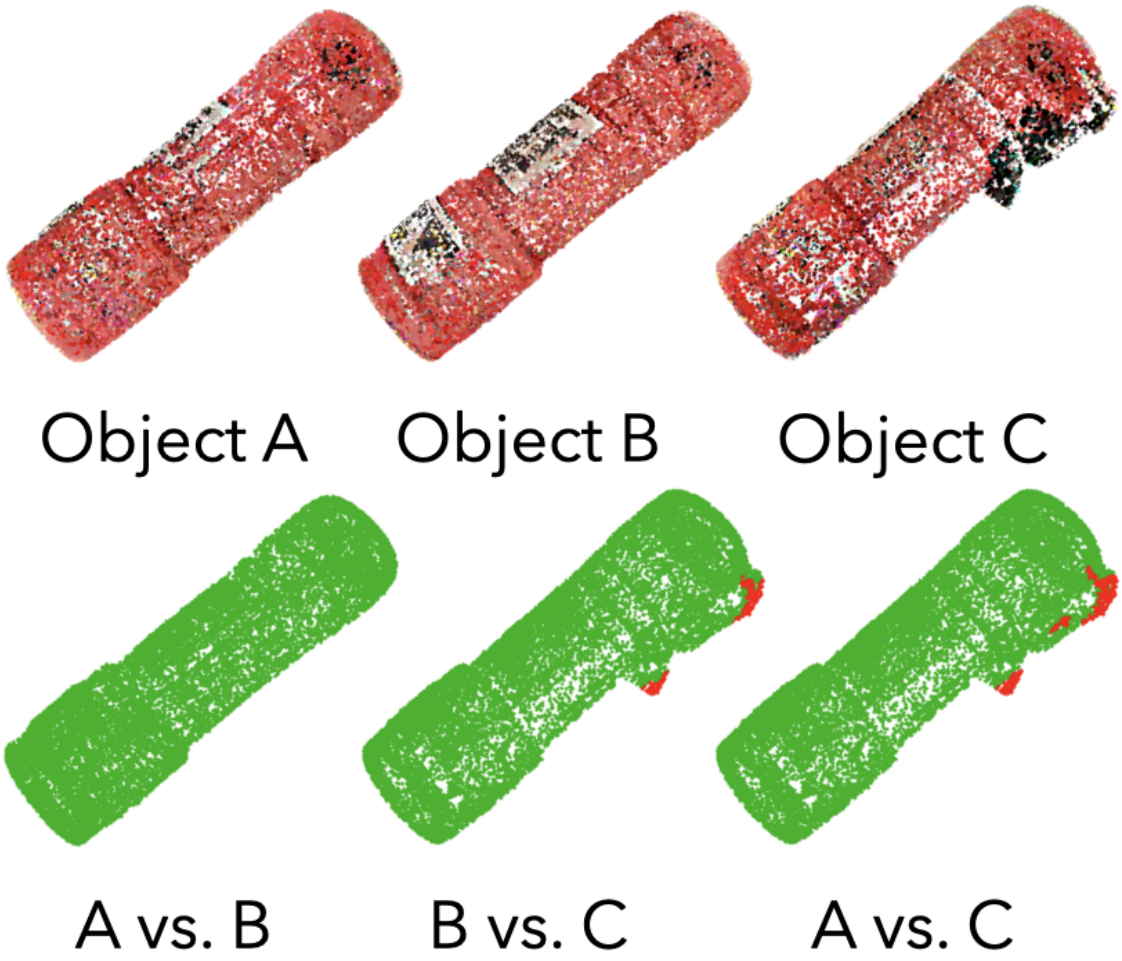

Figure 3.7: **Geometric Defect Detection** *Top Row:* The aligned point clouds of three scanned objects. *Bottom Row:* The point cloud difference between any two point clouds. Green points are points that are within the minimum distance to any other point on the other point cloud while red points are points which exceed this threshold and are classified as defect points.

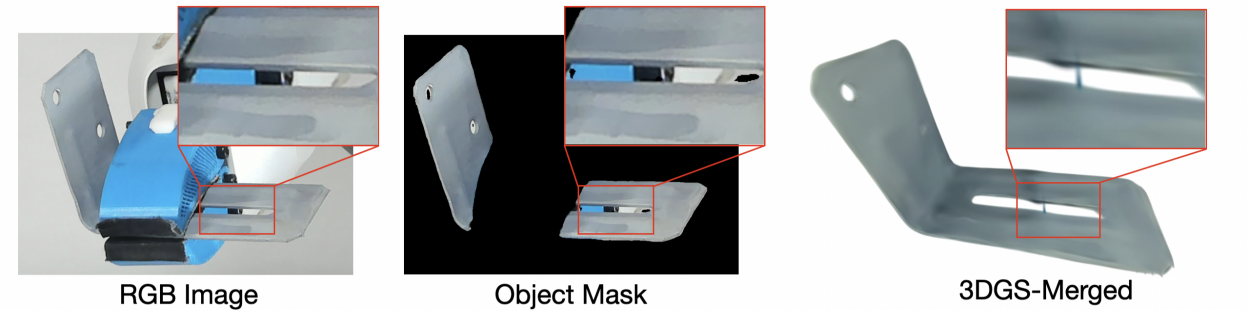

Figure 3.8: **Masking Failure Case** When objects contain cutouts, OMNI-SCAN may incorrectly include the pixels inside the cutout as part of the mask. This can lead to artifacts in the reconstruction, as shown in the figure, where the middle of the groove appears closed. from any other point exceeds our distance threshold of 4.5mm (empirically determined based on our set of objects to cause no false positives), then we classify it as a defective point.

Geometric Defect Detection Results OMNI-SCAN is able to correctly identify the geometric defect in 6 out of 7 trials. These results indicate that the point clouds generated by training a OMNI-SCAN are quite consistent among different undamaged objects as they have next to no defect points which exceeded our distance threshold of 4.5mm. This also further reinforces the ability of the alignment pipeline to properly align these models. Point cloud differencing fails on the pressure sensor with a geometric defect of slight sanding on one end of the object and a cut made on another end. These defects are marginal and the resulting point cloud does not noticeably differ from the two reference object point clouds.

3.5 Limitations

One limitation of OMNI-SCAN is with specularities. When scanning metallic objects, the color as well as the brightness can change depending on the pose of the camera to the object. This leads to issues with alignment and pixel differencing, since the same point on the object may look very different to the model depending on how it was grasped/ scanned. The system also relies on the handover pose as a good initialization for the Iterative Closest Point to estimate the transform between the left and right datasets. If the object slips significantly during handover, the resulting pose estimation ceases to be accurate, and the overall model quality suffers as a result. Since 3DGS models can contain gaussians in their interior, geometric differencing sometimes presents spurious false positives. Future work will explore mesh-based approaches for geometric differencing which better localize geometric defects.

3.6 Conclusion

We present Omni-Scan, a system for autonomous high-quality robotic creation of omni-directional digital twins and defect inspection. Experiments suggest that Omni-Scan constructs models with sufficient visual fidelity to detect visual and geometric defects on household, office, and industrial objects with up to 83% accuracy.

Chapter 4

MonoDuo: Using One Robot Arm to Learn Bimanual Robot Policies

4.1 Introduction

Bimanual robotic systems offer the potential to perform complex, coordinated manipulation tasks that are difficult or impossible for single-arm robots to execute. Many industrial and home tasks require two arms working in concert, with precise timing, spatial awareness, and physical coordination. However, a majority of available datasets and research infrastructure uses single-arm robots. This creates a bottleneck for learning bimanual policies, where the scarcity of bimanual robots significantly limits scalability.

We address this gap by proposing *MonoDuo*, a novel framework that enables learning bimanual manipulation policies using only single-arm robot demonstrations paired with human collaboration. MonoDuo builds on recent advances in cross-embodiment learning—techniques for transferring behaviors across different robot morphologies—and extends them to the challenging setting of single-arm to bimanual transfer. Specifically, MonoDuo begins with a human teleoperating a single-arm robot to perform one side of a bimanual task, while coordinating with a second human arm. The left-right roles are alternated across episodes, such that the human and the robot are equally included in both sides of the bimanual task, producing a balanced dataset for learning bimanual coordination. This dataset is then augmented into synthetic robot demonstrations generated for specified bimanual robot hardware, using state-of-the-art hand pose estimation, image and point cloud segmentation, and inpainting techniques. These synthetic demonstrations, grounded in real robot kinematics, are used to train bimanual manipulation policies.

We evaluate MonoDuo on a suite of 5 bimanual coordination tasks, including lifting a box with two arms, packing a backpack, zipping up a jacket, performing an object handover, and folding a piece of cloth. MonoDuo is capable of achieving zero-shot success using only data from a single-arm robot paired with a human, with success rates ranging from 35% to 70% on these tasks. We additional study a practical few-shot learning scenario, where only a

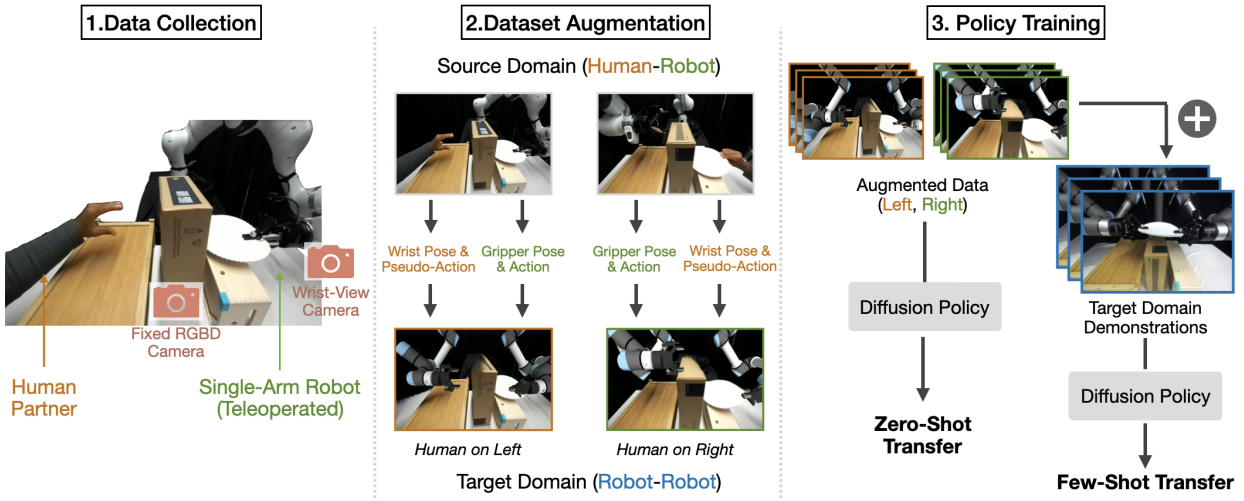

Figure 4.1: **Overview of MonoDuo.** We begin by teleoperating a single-arm robot to collaborate with a human on a bimanual task, alternating roles across episodes. This results in complementary interaction data covering both sides of the task. These human-robot demonstrations are then augmented into synthetic robot-robot demonstrations using segmentation and inpainting techniques, creating a visually and physically grounded dataset for bimanual robots. We train policies on this combined data—comprising real robot actions and human hand motions—enabling the model to learn coordinated bimanual behaviors.

small number of demonstrations on the target bimanual robot are available. In this setting, we show that MonoDuo improves sample efficiency significantly, increasing success rates by 65∼70% compared to policy without MonoDuo. This chapter makes four contributions:

1. **MonoDuo**, a novel framework for collecting demonstration data using one robot arm in collaboration with a human, synthesizing bimanual demonstrations, and learning bimanual manipulation policies when only a single-arm robot is available.

2. A data transformation pipeline that combines hand pose estimation, image and point cloud segmentation, and inpainting techniques to transform demonstration data collected with a single-arm robot and a human into bimanual demonstrations tailored to a specified bimanual robot.

3. Experiments suggesting that policies trained with MonoDuo can generalize zero-shot to previously unseen bimanual robot configurations, evaluated on a set of 5 bimanual tasks.

4. Experiments suggesting that MonoDuo significantly improves sample efficiency when finetuned with 25 bimanual robot demonstrations.

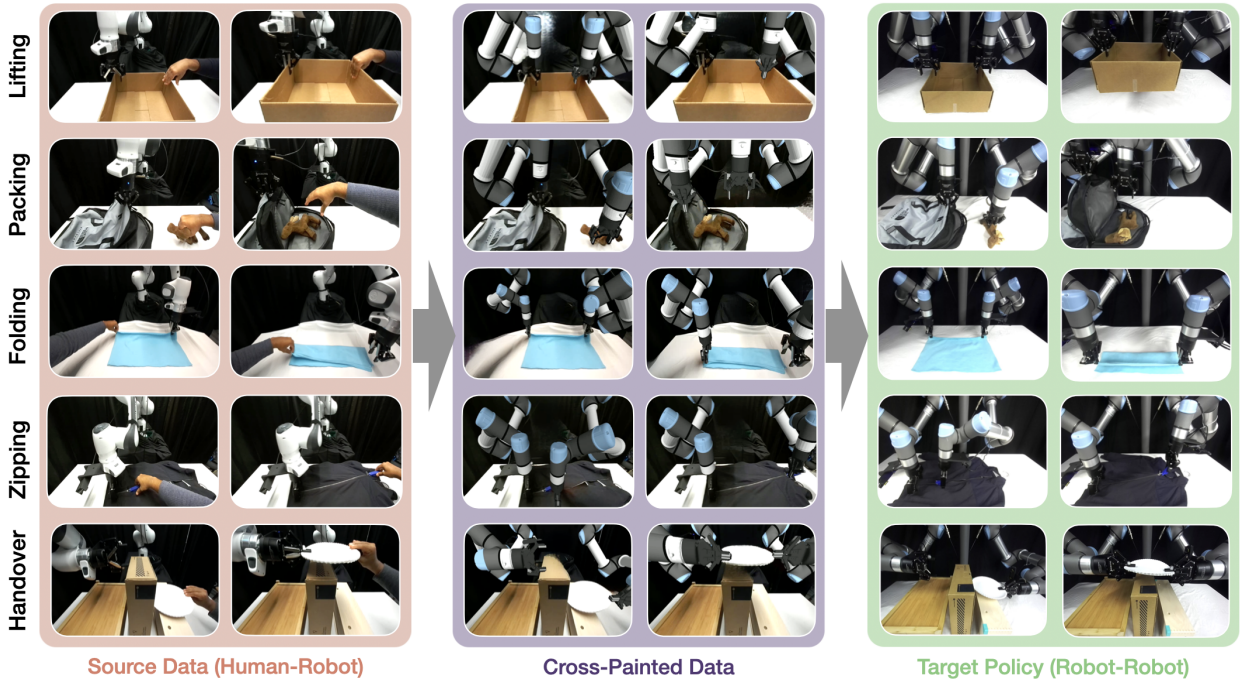

Figure 4.2: **From Human-Robot Demonstrations to Robot-Robot Policies.** Given collaborative demonstration trajectories between a single-arm robot and a human, MonoDuo uses state-of-the-art diffusion models to augment the image data and generate synthetic dataset tailored to a specified bimanual robot. Policies trained with the augmented dataset can be deployed on this target bimanual robot zero-shot. The same dataset can also be used to improve sample efficiency for few-shot learning.

4.2 Problem Statement

As described in Figure 4.1, MonoDuo collects a demonstration dataset $\mathcal{D}^{\mathcal{S}} = \{\tau_1^{\mathcal{S}}, \tau_2^{\mathcal{S}}, ..., \tau_n^{\mathcal{S}}\}$ consisting of $2N$ successful trajectories of a source robot-human pair $\mathcal{S} = (\mathcal{S}_r, \mathcal{S}_h)$ performing some task. Each trajectory $\tau_i^{\mathcal{S}} = (\{o_{1..H_i}^{\mathcal{S}}\}, \{p_{1..H_i}^{\mathcal{S}_r}\}, \{p_{1..H_i}^{\mathcal{S}_h}\}, \{a_{1..H_i}^{\mathcal{S}_r}\}, \{a_{1..H_i}^{\mathcal{S}_h}\})$, where $\{o_1^{\mathcal{S}}, ..., o_{H_i}^{\mathcal{S}}\}$ is a sequence of RGB-D camera observations, $\{p_1^{\mathcal{S}_r}, ..., p_{H_i}^{\mathcal{S}_r}\}$ is the sequence of corresponding robot state observations, $\{p_1^{\mathcal{S}_h}, ..., p_{H_i}^{\mathcal{S}_h}\}$ is the sequence of corresponding human hand state observations, $\{a_1^{\mathcal{S}_r}, ..., a_{H_i}^{\mathcal{S}_r}\}$ is the sequence of corresponding robot actions, and $\{a_1^{\mathcal{S}_r}, ..., a_{H_i}^{\mathcal{S}_r}\}$ is the sequence of corresponding human pseudo-actions. The robot state observations consist of current gripper pose and opening width. The human hand state observations consist of parameters returned from a hand state estimation algorithm. Each robot action or human pseudo-actions consists of gripper pose and opening width. Since human hand has a much different morphology from parallel-jaw gripper, MonoDuo includes a module to translate estimated human hand pose to gripper pose and opening width. We will elaborate more on how the human pseudo-actions are obtained from estimated hand

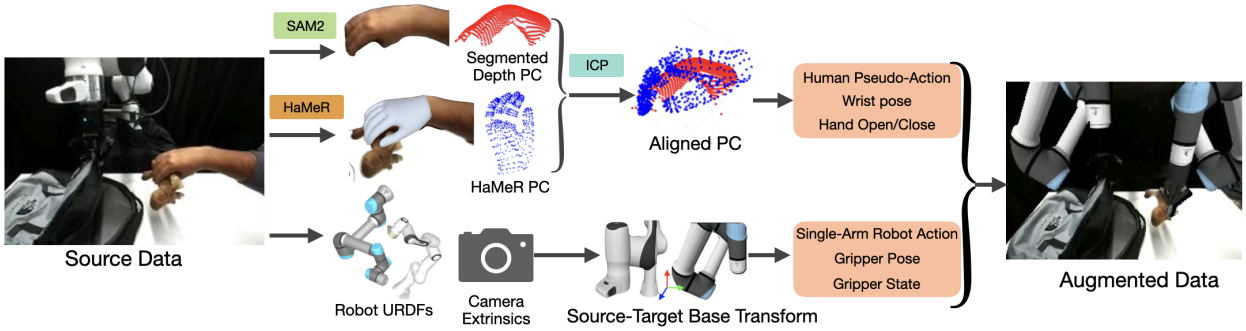

Figure 4.3: **Data Collection and Dataset Augmentation.** *Left*: We apply HaMeR [1] to estimate the hand pose at each frame and refine with ICP [2, 3]. The refined hand pose is then converted into pseudo-actions in the source dataset. *Right*: We perform cross-painting from both the source robot and the human arm to the target robot.

states in Section 4.3.

MonoDuo then augments $\mathcal{D}^{\mathcal{S}}$ into $\mathcal{D}^{\text{Aug}}$ to train a bimanual robot policy that can be deployed on a specified target bimanual robot $\mathcal{T}$ without test-time modification. This is illustrated in Figure 4.2. We assume the grippers of robot $\mathcal{S}$ and robot $\mathcal{T}$ are both parallel-jaw grippers, and that each single-arm robot with gripper has kinematics that can be approximated with a human arm and hand. We also assume fixed and known camera poses for both the source and target domains. This allows us to render robots with known URDFs in ways that are within the training image distribution. Similar to prior work [100, 88, 126, 127], we use Cartesian control and assume known inverse kinematics of the end-effector coordinate frames with respect to robot bases, such that we can use a rigid transformation $T_{\mathcal{T}}^{\mathcal{S}}$ to preprocess the data and align all end-effector poses $p^{\mathcal{S}} = T_{\mathcal{T}}^{\mathcal{S}} p^{\mathcal{T}}$ and actions $a^{\mathcal{S}} = T_{\mathcal{T}}^{\mathcal{S}} a^{\mathcal{T}}$ into the same vector space. Thus, for notational convenience, we omit the superscript differentiating end-effector poses and actions between $\mathcal{S}$ and $\mathcal{T}$.

After data augmentation, we learn a policy $\pi(a_t | o_t^{\mathcal{T}}, p_t)$ on $\mathcal{D}^{\text{Aug}}$ using a behavior cloning algorithm of choice. At test time, this policy takes as inputs the observations from the target robot and outputs actions that can be deployed on the target robot. In a second set of experiments, we co-train $\mathcal{D}^{\text{Aug}}$ with a small number of demonstration data $\mathcal{D}^{\mathcal{T}}$ directly obtained from the target bimanul robot, and study how this leads to improvement on few-shot generalization.

4.3 MonoDuo

In this section, we describe more details of how MonoDuo enables the learning of bimanual robot policies when only a single-arm robot is available. An overview is shown in Figure 4.1.

Data Collection

For each bimanual task, we collect a source dataset $\mathcal{D}^{\mathcal{S}}$ using a human to teleoperate a single-arm robot to collaborate with a human partner on the task. To ensure a balanced data distribution, the roles of left-arm and right-arm are alternated across episodes. Specifically, human arm and robot collect N trajectories on each side, where the total number of trajectories is $2N$. Data is collected in the format outlined in Section 4.2.

We resolve the morphology gap between human and robot by translating the human arm-hand motions into robot-like "pseudo-actions." This is feasible based on two observations: (1) human wrist pose can be approximated as robot end-effector pose; (2) human hand pose can be approximated as gripper state. We begin by estimating the 3D human hand pose at each timestep—specifically, by applying HaMeR [1] to each RGB image from camera observation $o_t^{\mathcal{S}}$. HaMeR predicts 21 keypoints, $\hat{\mathbf{X}}_t \in \mathbb{R}^{21\times 3}$, corresponding to anatomical landmarks following the MANO [128] model. Since HaMeR struggles to estimate the absolute 3D pose due to its reliance on a monocular image, we incorporate depth to refine this estimate. In the RGB image observation, we use SAM2 [129] to obtain a segmentation mask of the hand; then, we extract a partial point cloud of the hand by applying the segmentation mask on the aligned depth image. Next, we align the HaMeR-predicted mesh $\hat{\mathbf{V}}_t$ with the segmented hand point cloud $\mathbf{P}_t$ via Iterative Closest Point (ICP) registration [2], obtaining the optimal rigid transformation $\mathbf{T}_t \in SE(3)$ such that $\mathbf{P}_t \approx \mathbf{V}_t = \mathbf{T}_t\hat{\mathbf{V}}_t$. Since $\hat{\mathbf{V}}_t$ and $\hat{\mathbf{X}}_t$ are internally consistent, we can apply $\mathbf{T}_t$ to the predicted keypoints to refine their positions: $\mathbf{X}_t = \mathbf{T}_t\hat{\mathbf{X}}_t$. Once the keypoints $\mathbf{X}_t$ are refined, we define the pseudo-actions $a_t^{\mathcal{S}_h}$ in $\mathcal{D}^{\mathcal{S}}$ as follows: the end-effector pose is set as the estimated wrist pose, and the gripper opening is computed as a binary variable based on the scalar angle defined by three MANO [128] landmarks: thumb fingertip, index finger fingertip, and index proximal frame. We set a threshold value for the scalar angle value, such that angle below which is translated to a closed gripper.

Dataset Augmentation

Given the source dataset $\mathcal{D}^{\mathcal{S}}$, we aim to augment it into $\mathcal{D}^{\text{Aug}}$ to learn a bimanual policy that can be deployed on the target bimanual robot $\mathcal{T}$. To this end, we apply "cross-painting"—which in prior works [99, 100] means replacing the source robot with the target robot in the camera observations at test time so that it appears to the policy as if the source robot were performing the task. In this work, we extend cross-painting to also include human as a data source. We describe the details below, and illustrate the cross-painting procedure in Figure 4.2.

Source Robot to Target Robot Cross-Painting. We leverage knowledge of the source and target robot URDFs and camera poses to perform robot-robot cross-painting at training time, as illustrated in Figure 4.3. First, given known camera extrinsics, we re-project the images from the source domain to the target domain given that depth sensing is available.

Next, given the RGB image observation and joint angles of the source robot, we use a renderer to determine which image pixels correspond to the source robot and mask out these pixels. Then, we inpaint the missing pixels using a video inpainting model E^2FGVI [130]. Finally, we use the URDF of the target robot to solve for the joint angles that would put its end effector at the same pose as that of the source robot, render it using a simulator, and overlay it onto the source image. For the gripper, we similarly compute and set the joints of the target robot gripper in the renderer so that its width would roughly match that of the source robot's gripper. To prevent the trained policy on the augmented data from overfitting to the synthetic robot visuals, we perform random brightness augmentation to the generated robot before pasting it. Previous work [99] has demonstrated this random brightness augmentation to significantly help improve the performance of trained policies.

Human to Target Robot Cross-Painting. Cross-painting from human to target robot largely follows the same process as robot-robot cross-painting, except that we segment out the pixels corresponding to the human arm using SAM2 [129] before replacing the human embodiment with a robot. The target model is similarly rendered, with its end effector pose and gripper opening width corresponding to the pseudo-action extracted.

Policy Training

After applying dataset augmentation, we can train a policy π based on the Diffusion Policy architecture [64] on the augmented dataset $\mathcal{D}^{Aug}$ and zero-shot deploy the policy on the target robot $\mathcal{T}$. The policy input is RGB image observations and bimanual robot state observations; policy output is bimanual robot actions. For challenging tasks or when there is a large difference in the dynamics between the robots, we can also collect a small demonstration dataset $\mathcal{D}^{\mathcal{T}}$ on the target robot directly and few-shot finetune π on $\mathcal{D}^{\mathcal{T}}$ to further improve policy performance.

4.4 Experiments

Hardware Setup and Task Definition

We use a Franka arm as the single-arm source robot, and a pair of UR5e arms setup as the bimanual target robot. For RGB-D data collection, we use a ZED2 stationary fixed camera and a ZED-mini wrist-mounted camera. We design five bimanual tasks for policy evaluation: (1) **Box Lifting**: the robot needs to coordinate the two grippers to lift up a box; (2) **Backpack Packing**: the robot needs to use one gripper to open a backpack, pick up a toy using the other gripper, put the toy into the backpack, and finally close the backpack with the first arm; (3) **Jacket Zipping**: the robot uses one arm to pin the jacket and the other arm to grasp and zip up the zipper of a jacket; (4) **Plate Handover**: the robot uses one gripper to pick up a plate and hands it over to the other gripper, while the other gripper

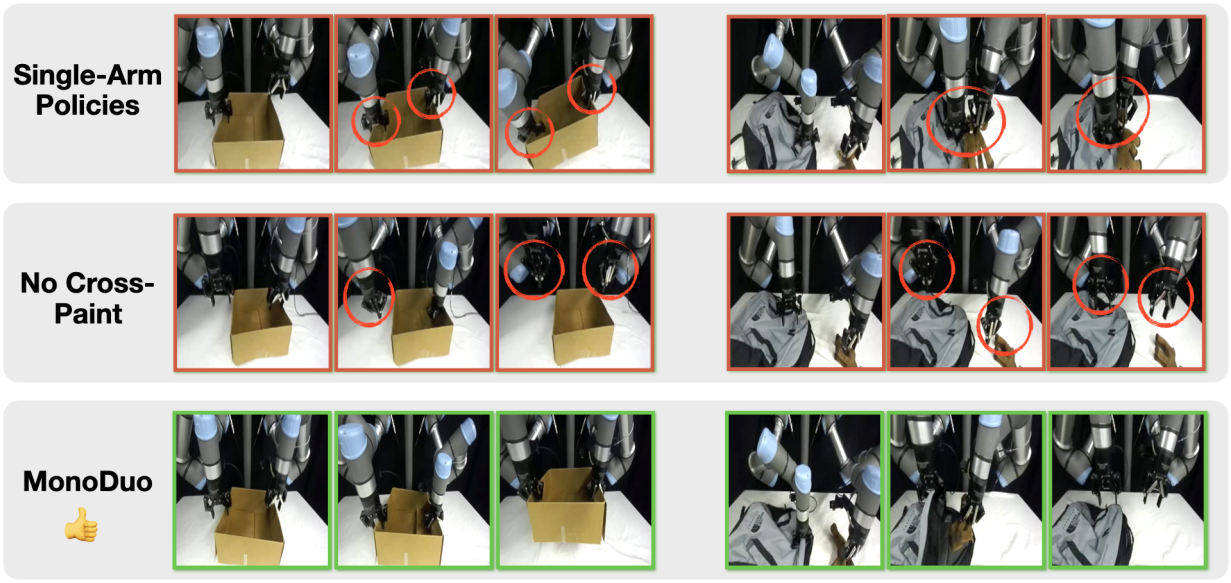

Figure 4.4: **Examples of zero-shot rollout on the target bimanual UR5e.** *Left:* Lift Box; *Right:* Pack Bag. Single-Arm policies do not coordinate the actions well, leading to asynchronous movements as shown in the Lift Box task and collision in the Pack Bag task. Policies trained without cross-painting is less robust and misgrasps often. MonoDuo exhibits coordinated behaviors while being precise.

needs to come to the waiting pose, stably grasp the plate, and put it down; (5) **Cloth Folding**: the robot needs to coordinate the two grippers to fold a piece of cloth by half. All tasks require highly coordinated behaviors of two arms and cannot be accomplished with a single-arm robot.

Implementation Details

We collect 200 demonstration trajectories on the source robot for each task, half of which has human on the left side and the other half the right side. For generalization experiment, we additionally collect 25 trajectories on the target robot. On both setups, we ues Meta Quest as the teleoperation device. We use the UNet-based Diffusion Policy as outlined in Chi et al. [64], with a ResNet encoder for visual observations. Policies take Cartesian proprioception, 2 image observations, and predict Cartesian end-effector actions.

Results

Zero-Shot Bimanual Policies. We report success rates of zero-shot deployed MonoDuo policies on each evaluation task in Table 4.1 and visualize their qualitative behaviors in

Policies	Policy Attributes				Task Success Rates				
	Use Robot	Cross Paint	Pseudo Action	Weight Sharing	Lift Box	Pack Bag	Zip Jacket	Handover Plate	Fold Cloth
Single-Arm Policies (Naive)	✓	✓			15%	10%	15%	0%	5%
Pure Human Videos		✓	✓	✓	10%	0%	0%	0%	0%
Ablation: No Cross-Paint	✓		✓	✓	40%	30%	15%	10%	15%
Ablation: No Pseudo-Action	✓	✓		✓	50%	20%	20%	30%	25%
MonoDuo	✓	✓	✓	✓	**70%**	**55%**	**45%**	**35%**	**35%**

Table 4.1: **Zero-shot experiments comparing MonoDuo with baselines.** Each policy is evaluated on five manipulation tasks in a zero-shot transfer setting from Franka-human demos to a bimanual UR5e.

Figure 4.4. These results show that MonoDuo is able to effectively bridge both the visual and physical gaps among different robots and human, allowing one to learn bimanual policies when only a single-arm robot is available.

Few-Shot MonoDuo. We study the finetuned performance of MonoDuo by training policies with an addition of 25 trajectories collected obtained from direct teleoperation on the target robot. This corresponds to a common realistic scenario, where only a small number of demonstrations on the target bimanual robot is available. Results in Table 4.3 show that MonoDuo improve learning efficiency significantly, reaching a much higher performance level with the same number of real demonstrations. Notably, few-shot MonoDuo is able to increase the success rate of *box lifting* task from 30% to 100%, *backpack packing* from 25% to 90%, and *jacket zipping* from 5% to 75%. These results highlight how MonoDuo can greatly complement limited real-world bimanual data.

Comparison with Baselines. We evaluate MonoDuo against 4 baselines: (1) *Single-Arm Policies*: Two independent single-arm policies, each conditioned on the cross-painted global observation and its respective robot state, trained to predict the action for a single arm. (2) *Pure Human Videos*: A policy trained solely on bimanual human-only demonstrations, using cross-painting to simulate robot embodiment and pose estimation of both hands. (3) *No Cross-Paint*: An ablation that removes the visual domain alignment step, training instead on raw images while still leveraging both human and robot action supervision. (4) *No Pseudo-Action*Similar to MonoDuo in using a unified policy architecture, but excludes human pseudo-actions during training, relying only on robot action supervision. Quantitative results in Table 4.1 show that all three core components of MonoDuo —joint human-robot demonstrations, robot-robot cross-painting, and human-robot cross-painting—are essential for learning effective bimanual coordination policies. Figure 4.4 show some qualitative examples, and we analyze key insights from ablation studies below.

Importance of Weight-Sharing. Our results indicate that using two disjoint single-arm policies, even when paired with cross-painted visual input, fails to produce reliable coordination. Each arm tends to execute its part of the task independently, leading to asynchronous behavior. While some tasks may succeed occasionally, the overall quality is poor—for example, lifting a box unevenly—and no success is observed in tasks demanding precise temporal synchronization.

Importance of Cross-Painting. Removing cross-painting results in a 20–30% drop in success across all tasks, demonstrating the critical role of visual domain alignment in enabling effective transfer. Cross-painting helps mitigate the embodiment gap and enables the model to generalize better.

Value of Human Pseudo-Actions. Training solely on human video data yields nearly zero success in zero-shot settings, primarily due to noisy hand pose estimations. To enable fair comparison, we also include results of MonoDuo trained without wrist cameras in Table 4.2. Interestingly, we observe that incorporating human pseudo-actions alongside robot actions improves performance, especially in tasks where hand pose estimation is more accurate. Compared to using human videos alone, leveraging both human and robot actions reduces the dependency on high-fidelity human pose estimation while enhancing policy performance.

	No WristCam	*With WristCam*
Lift Box	60%	70%
Pack Bag	40%	55%
Zip Jacket	25%	45%

Table 4.2: **Impact of Wrist Camera on Zero-Shot Performance.** Using only a third-person camera yields strong results, but wrist-mounted cameras improve precision in tasks requiring fine manipulation, such as zipper grasping.

	Scratch	*Few-Shot MonoDuo*
Lift Box	30%	100%
Pack Bag	25%	90%
Zip Jacket	5%	75%

Table 4.3: **Few-Shot Learning with MonoDuo.** Incorporating 25 target robot demonstrations enables MonoDuo to significantly outperform policies trained from scratch, demonstrating improved sample efficiency.

4.5 Conclusion

We present MonoDuo, a novel framework for learning bimanual robot policies using only demonstrations from a single-arm robot in collaboration with a human. By alternating roles between human and robot across episodes and applying vision-based augmentation techniques, MonoDuo generates synthetic bimanual demonstrations tailored to a specified target robot. This approach enables training policies that generalizes zero-shot to previously unseen bimanual robot configurations, and significantly improves sample efficiency in low-data

regimes. We validate MonoDuo on five challenging bimanual manipulation tasks, demonstrating its effectiveness and superior performance over baselines. We believe MonoDuo can be a scalable and accessible solution for bimanual robot learning.

4.6 Limitations and Future Work

While MonoDuo introduces a scalable framework for learning bimanual manipulation policies using only single-arm robot data, several limitations remain. First, the approach assumes fixed and known camera calibration for both source and target domains. This assumption simplifies rendering and cross-painting but limits applicability in environments where such calibration is not readily available. Second, MonoDuo requires depth sensing to refine 3D hand pose estimates and perform accurate segmentation and augmentation. As a result, it requires an RGB-D camera.

In addition, we do not explicitly tackle generalization across novel backgrounds or camera viewpoints, which often occur in real-world settings. Future work could combine MonoDuo with prior orthogonal approaches such as object, background, camera, and task augmentation [99, 131, 132]. Finally, all experiments in this chapter use parallel-jaw grippers. Extending MonoDuo to handle dexterous hands requires fine-grained finger motion tracking, and would be a promising future work.

Chapter 5

Conclusion

This thesis explored how bimanual robots can serve not only as execution agents but also as scalable platforms for data generation. We investigated this through two directions: Omni-Scan, a system for generating high-fidelity 3D object models using a bimanual scanning setup, and MonoDuo, a framework for learning transferable manipulation policies from single-arm demonstrations. As dual-arm systems become increasingly accessible and capable, we hope that the contributions of Omni-Scan and MonoDuo will advance the field of bimanual dexterity and enable new lines of research beyond traditional execution-focused applications.

Omni-Scan introduces a novel approach to in-hand scanning that turns a bimanual robot into an active, dynamic viewpoint controller for 3D reconstruction. Through coordinated handovers between two grippers, the system reveals previously occluded object surfaces without relying on multi-camera arrays or wrist-mounted sensors. By leveraging vision models such as DepthAnything, Segment Anything, and RAFT, the pipeline effectively segments the object from both the robot and the background. This enables the training of high-fidelity 3D Gaussian Splat models that are not only visually realistic but also well-suited for downstream applications such as part inspection and defect detection. Applied to 12 industrial and household objects, Omni-Scan achieved an average defect detection accuracy of 83%, demonstrating its practical utility.

MonoDuo continues the theme of scalable data generation by addressing the data scarcity challenge in bimanual policy learning. Rather than relying on costly and difficult-to-obtain dual-arm demonstrations, it synthesizes them from alternating single-arm teleoperation episodes involving human-robot collaboration. By combining vision-based tracking with kinematic grounding, it generates bimanual demonstrations suitable for training policies that generalize across robot embodiments. Experiments across five diverse dual-arm tasks show that policies trained without any real bimanual demonstrations can generalize zero-shot to new dual-arm configurations. Furthermore, ,few-shot finetuning MonoDuo with just 25 real demonstrations on the target robot leads to substantial performance improvements, making it a practical and sample-efficient approach for learning bimanual manipulation skills.

Despite these contributions, both systems face several limitations. Most notably, they

both rely on accurate camera calibration and reliable depth sensing, which can limit robustness in unstructured or dynamically lit environments. Omni-Scan is particularly sensitive to lighting conditions and surface properties—specular or reflective objects can disrupt view consistency and disrupt alignment. The system also depends on stable handovers; slippage during re-grasping can introduce errors in point cloud registration, degrading the final object reconstruction quality. MonoDuo, on the other hand, assumes consistent backgrounds and fixed camera viewpoints. It does not explicitly address generalization to varying visual conditions, which are common in real-world deployment. Addressing these limitations will be key to improving the robustness and adaptability of both systems.

Future work can extend these systems along several directions. For Omni-Scan, producing mesh and texture outputs—similar to the approach used in [133]—may improve reconstruction quality and surface fidelity compared to Gaussian Splat models. Replacing ICP-based alignment with learning-based registration methods could also enhance robustness to imperfect handovers and minor slippage. For MonoDuo, integrating recent advances in sim-to-real transfer and domain randomization could improve policy generalization across environments. It may also be valuable to combine MonoDuo with orthogonal strategies such as object, background, camera, and task augmentation [99, 131, 132]. Finally, extending MonoDuo to support dexterous hands rather than parallel-jaw grippers would require fine-grained finger motion tracking, but presents an exciting direction for enabling more complex manipulation skills.

Bibliography

[1] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

[2] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992.

[3] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992.

[4] Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin

Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models, 2025.

[5] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho

Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. 2024.

[6] Andrew Goldberg, Kavish Kondap, Tianshuang Qiu, Zehan Ma, Letian Fu, Justin Kerr, Huang Huang, Kaiyuan Chen, Kuan Fang, and Ken Goldberg. Blox-net: Generative design-for-robot-assembly using vlm supervision, physics, simulation, and a robot with reset.

[7] Simeon* Adebola, Tara* Sadjadpour, Karim* El-Refai, Will Panitch, Zehan Ma, Roy Lin, Tianshuang Qiu, Shreya Ganti, Charlotte Le, Jaimyn Drake, and Ken Goldberg. Automating deformable gasket assembly. 2024.

[8] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

[9] Michal Adamkiewicz, Timothy Chen, Adam Caccavale, Rachel Gardner, Preston Culbertson, Jeannette Bohg, and Mac Schwager. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*, 7(2):4606–4613, 2022.

[10] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.

[11] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

[12] Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12861–12870, 2022.

[13] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio:

A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–12, 2023.

[14] Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, and Wenping Wang. F2-nerf: Fast neural radiance field training with free camera trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4150–4159, 2023.

[15] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023.

[16] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.

[17] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 333–350. Springer, 2022.

[18] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023.

[19] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5501–5510, 2022.

[20] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021.

[21] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4284, 2023.

[22] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[23] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for

slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12786–12796, June 2022.

[24] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J. Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6229–6238, October 2021.

[25] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3437–3444. IEEE, 2023.

[26] Yunzhu Li, Shuang Li, Vincent Sitzmann, Pulkit Agrawal, and Antonio Torralba. 3d neural scene representations for visuomotor control. In *Conference on Robot Learning*, pages 112–123. PMLR, 2022.

[27] Danny Driess, Ingmar Schubert, Pete Florence, Yunzhu Li, and Marc Toussaint. Reinforcement learning with neural radiance fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[28] Justin Kerr, Letian Fu, Huang Huang, Yahav Avigal, Matthew Tancik, Jeffrey Ichnowski, Angjoo Kanazawa, and Ken Goldberg. Evo-neRF: Evolving neRF for sequential robot grasping of transparent objects. In *6th Annual Conference on Robot Learning*, 2022.

[29] Jeffrey Ichnowski*, Yahav Avigal*, Justin Kerr, and Ken Goldberg. Dex-NeRF: Using a neural radiance field to grasp transparent objects. In *Conference on Robot Learning (CoRL)*, 2020.

[30] Adam Rashid, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Yunliang Chen, Angjoo Kanazawa, and Ken Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In *Conference on Robot Learning*, 2023.

[31] Justin Kerr, Chung Min Kim, Mingxuan Wu, Brent Yi, Qianqian Wang, Ken Goldberg, and Angjoo Kanazawa. Robot see robot do: Imitating articulated object manipulation with monocular 4d reconstruction. In *8th Annual Conference on Robot Learning*, 2024.

[32] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. In *7th Annual Conference on Robot Learning*, 2023.

[33] Arunkumar Byravan, Jan Humplik, Leonard Hasenclever, Arthur Brussee, Francesco Nori, Tuomas Haarnoja, Ben Moran, Steven Bohez, Fereshteh Sadeghi, Bojan Vujatovic, et al. Nerf2real: Sim2real transfer of vision-guided bipedal motion skills using neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9362–9369. IEEE, 2023.

[34] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–16, 2023.

[35] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2022.

[36] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pages 1–16, 2016.

[37] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023.

[38] Licheng Zhong, Lixin Yang, Kailin Li, Haoyu Zhen, Mei Han, and Cewu Lu. Colorneus: Reconstructing neural implicit surfaces with color. In *2024 International Conference on 3D Vision (3DV)*, pages 631–640. IEEE, 2024.

[39] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. BundleSDF: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.

[40] Aamir Khan, Carmelo Mineo, Gordon Dobie, Charles Macleod, and Gareth Pierce. Vision guided robotic inspection for parts in manufacturing and remanufacturing industry. *Journal of Remanufacturing*, 11(1):49–70, 2021.

[41] Omid Davtalab, Ali Kazemian, Xiao Yuan, and Behrokh Khoshnevis. Automated inspection in robotic additive manufacturing using deep learning for layer deformation detection. *Journal of Intelligent Manufacturing*, 33(3):771–784, 2022.

[42] Simon Stepputtis, Maryam Bandari, Stefan Schaal, and Heni Ben Amor. A system for imitation learning of contact-rich bimanual manipulation policies. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11810–11817. IEEE, 2022.

[43] Jennifer Grannen, Yilin Wu, Brandon Vu, and Dorsa Sadigh. Stabilize to act: Learning to coordinate for bimanual manipulation. In *Conference on Robot Learning*, pages 563–576. PMLR, 2023.

[44] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *RSS*, 2023.

[45] Hongjie Fang, Hao-Shu Fang, Yiming Wang, Jieji Ren, Jingjing Chen, Ruo Zhang, Weiming Wang, and Cewu Lu. Low-cost exoskeletons for learning whole-arm manipulation in the wild. In *ICRA*, 2023.

[46] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024.

[47] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163. IEEE, 2024.

[48] Aadhithya Iyer, Zhuoran Peng, Yinlong Dai, Irmak Guzey, Siddhant Haldar, Soumith Chintala, and Lerrel Pinto. Open teach: A versatile teleoperation system for robotic manipulation. *arXiv preprint arXiv:2403.07870*, 2024.

[49] Toru Lin, Yu Zhang, Qiyang Li, Haozhi Qi, Brent Yi, Sergey Levine, and Jitendra Malik. Learning visuotactile skills with two multifingered hands. *arXiv:2404.16823*, 2024.

[50] Binghao Huang, Yuanpei Chen, Tianyu Wang, Yuzhe Qin, Yaodong Yang, Nikolay Atanasov, and Xiaolong Wang. Dynamic handover: Throw and catch with bimanual hands. *arXiv preprint arXiv:2309.05655*, 2023.

[51] Toru Lin, Zhao-Heng Yin, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Twisting lids off with two hands. *arXiv:2403.02338*, 2024.

[52] Toru Lin, Kartik Sachdev, Linxi Fan, Jitendra Malik, and Yuke Zhu. Sim-to-real reinforcement learning for vision-based dexterous manipulation on humanoids. *arXiv:2502.20396*, 2025.

[53] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834. IEEE, 2021.

[54] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

[55] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023.

[56] Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation. In *8th Annual Conference on Robot Learning*, 2024.

[57] Arpit Bahety, Priyanka Mandikal, Ben Abbatematteo, and Roberto Martín-Martín. Screwmimic: Bimanual imitation from human videos with screw space projection. *arXiv preprint arXiv:2405.03666*, 2024.

[58] Yifeng Zhu, Arisrei Lim, Peter Stone, and Yuke Zhu. Vision-based manipulation from single human video with open-world object graphs. *arXiv preprint arXiv:2405.20321*, 2024.

[59] Huayi Zhou, Ruixiang Wang, Yunxin Tai, Yueci Deng, Guiliang Liu, and Kui Jia. You only teach once: Learn one-shot bimanual robotic manipulation from video demonstrations. *arXiv preprint arXiv:2501.14208*, 2025.

[60] Yuanpei Chen, Chen Wang, Yaodong Yang, and C Karen Liu. Object-centric dexterous manipulation from human motion data. *arXiv preprint arXiv:2411.04005*, 2024.

[61] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

[62] Maryam Zare, Parham M Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 2024.

[63] Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on robot learning*, pages 158–168. PMLR, 2022.

[64] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Learning agile robotic locomotion skills by imitating animals*, 2023.

[65] Open X-Embodiment Collaboration, Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, Antonin Raffin, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Brian Ichter, Cewu Lu, Charles Xu, Chelsea Finn, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Chuer Pan, Chuyuan Fu, Coline Devin, Danny Driess, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Federico Ceola, Fei Xia, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Giulio Schiavi, Hao Su, Hao-Shu Fang, Haochen Shi, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homer Walke, Hongjie Fang, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky

Liang, Jaehyung Kim, Jan Schneider, Jasmine Hsu, Jeannette Bohg, Jeffrey Bingham, Jiajun Wu, Jialin Wu, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jitendra Malik, Jonathan Tompson, Jonathan Yang, Joseph J. Lim, João Silvério, Junhyek Han, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Zhang, Keyvan Majd, Krishan Rana, Krishnan Srinivasan, Lawrence Yunliang Chen, Lerrel Pinto, Liam Tan, Lionel Ott, Lisa Lee, Masayoshi Tomizuka, Maximilian Du, Michael Ahn, Mingtong Zhang, Mingyu Ding, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Pannag R Sanketi, Paul Wohlhart, Peng Xu, Pierre Sermanet, Priya Sundaresan, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Sherry Moore, Shikhar Bahl, Shivin Dass, Shuran Song, Sichun Xu, Siddhant Haldar, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Sudeep Dasari, Suneel Belkhale, Takayuki Osa, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Vidhi Jain, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiaolong Wang, Xinghao Zhu, Xuanlin Li, Yao Lu, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yueh hua Wu, Yujin Tang, Yuke Zhu, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zhuo Xu, and Zichen Jeff Cui. Open X-Embodiment: Robotic learning datasets and RT-X models. IEEE International Conference on Robotics and Automation, 2024.

[66] Chen Yu, Weinan Zhang, Hang Lai, Zheng Tian, Laurent Kneip, and Jun Wang. Multi-embodiment legged robot control as a sequence modeling problem. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7250–7257. IEEE, 2023.

[67] Tao Chen, Adithyavairavan Murali, and Abhinav Gupta. Hardware conditioned policies for multi-robot transfer learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[68] Lin Shao, Fabio Ferreira, Mikael Jorda, Varun Nambiar, Jianlan Luo, Eugen Solowjow, Juan Aparicio Ojea, Oussama Khatib, and Jeannette Bohg. Unigrasp: Learning a unified model to grasp with multifingered robotic hands. *IEEE Robotics and Automation Letters*, 5(2):2286–2293, 2020.

[69] Zhenjia Xu, Beichun Qi, Shubham Agrawal, and Shuran Song. Adagrasp: Learning an adaptive gripper-aware grasping policy. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4620–4626. IEEE, 2021.

[70] Tingwu Wang, Renjie Liao, Jimmy Ba, and Sanja Fidler. Nervenet: Learning structured policy with graph neural networks. In *International conference on learning representations*, 2018.

[71] Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Martin Riedmiller, Raia Hadsell, and Peter Battaglia. Graph networks as learnable physics engines for inference and control. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4470–4479. PMLR, 10–15 Jul 2018.

[72] Deepak Pathak, Christopher Lu, Trevor Darrell, Phillip Isola, and Alexei A Efros. Learning to control self-assembling morphologies: a study of generalization via modularity. *Advances in Neural Information Processing Systems*, 32, 2019.

[73] Ashish Malik. Zero-shot generalization using cascaded system-representations. *arXiv preprint arXiv:1912.05501*, 2019.

[74] Wenlong Huang, Igor Mordatch, and Deepak Pathak. One policy to control them all: Shared modular policies for agent-agnostic control. In *International Conference on Machine Learning*, pages 4455–4464. PMLR, 2020.

[75] V Kurin, M Igl, T Rocktaschel, W Boehmer, and S Whiteson. My body is a cage: the role of morphology in graph- based incompatible control. In *Proceedings of the International Conference on Learning Representations*. OpenReview, 2021.

[76] Amaury Depierre, Emmanuel Dellandréa, and Liming Chen. Jacquard: A large scale dataset for robotic grasp detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3511–3516. IEEE, 2018.

[77] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, pages 651–673. PMLR, 2018.

[78] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436, 2018.

[79] Clemens Eppner, Arsalan Mousavian, and Dieter Fox. ACRONYM: A large-scale grasp dataset based on simulation. In *2021 IEEE Int. Conf. on Robotics and Automation, ICRA*, 2020.

[80] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home, 2023.

[81] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Junbo Wang, Haoyi Zhu, and Cewu Lu. RH20T: A robotic dataset for learning diverse skills in one-shot. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023.

[82] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. In *Robotics: Science and Systems (RSS) XVIII*, 2022.

[83] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[84] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-Z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning (CoRL)*, pages 991–1002, 2021.

[85] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. RT-1: Robotics transformer for real-world control at scale. *Robotics: Science and Systems (RSS)*, 2023.

[86] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[87] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: General robot manipulation with multimodal prompts. *International Conference on Machine Learning (ICML)*, 2023.

[88] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. GNM: A general navigation model to drive any robot. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7226–7233. IEEE, 2023.

[89] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. ViNT: A Foundation Model for Visual Navigation. In *7th Annual Conference on Robot Learning (CoRL)*, 2023.

[90] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.

[91] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[92] Austin Stone, Ted Xiao, Yao Lu, Keerthana Gopalakrishnan, Kuang-Huei Lee, Quan Vuong, Paul Wohlhart, Sean Kirmani, Brianna Zitkovich, Fei Xia, et al. Open-world object manipulation using pre-trained vision-language models. In *Conference on Robot Learning*, pages 3397–3417. PMLR, 2023.

[93] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

[94] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Transactions on Machine Learning Research*, 2022.

[95] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, 2022.

[96] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4788–4795. IEEE, 2024.

[97] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali-x: On scaling up a multilingual vision and language model, 2023.

[98] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, pages 8469–8488. PMLR, 2023.

[99] Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Muhammad Zubair Irshad, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. In *Conference on Robot Learning (CoRL)*, Munich, Germany, 2024.

[100] Lawrence Yunliang Chen, Kush Hari, Karthik Dharmarajan, Chenfeng Xu, Quan Vuong, and Ken Goldberg. Mirage: Cross-embodiment zero-shot policy transfer with cross-painting. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[101] Marion Lepert, Ria Doshi, and Jeannette Bohg. Shadow: Leveraging segmentation masks for zero-shot cross-embodiment policy transfer. In *Conference on Robot Learning (CoRL)*, Munich, Germany, 2024.

[102] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos, 2025.

[103] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video, 2024.

[104] Vidhi Jain, Maria Attarian, Nikhil J Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, et al. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers. *arXiv preprint arXiv:2403.12943*, 2024.

[105] Kushal Kedia, Prithwish Dan, Angela Chao, Maximus Adrian Pace, and Sanjiban Choudhury. One-shot imitation under mismatched execution, 2024.

[106] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. *arXiv preprint arXiv:2408.11812*, 2024.

[107] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[108] Guanxing Lu, Tengbo Yu, Haoyuan Deng, Season Si Chen, Yansong Tang, and Ziwei Wang. Anybimanual: Transferring unimanual policy for general bimanual manipulation. *arXiv preprint arXiv:2412.06779*, 2024.

[109] Masato Kobayashi, Jun Yamada, Masashi Hamaya, and Kazutoshi Tanaka. Lfdt: Learning dual-arm manipulation from demonstration translated from a human and robotic arm. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pages 1–8, 2023.

[110] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[111] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[112] Heng Yang, Jingnan Shi, and Luca Carlone. Teaser: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 37(2):314–333, 2021.

[113] Jiaqi Yang, Xiyu Zhang, Peng Wang, Yulan Guo, Kun Sun, Qiao Wu, Shikun Zhang, and Yanning Zhang. Mac: Maximal cliques for 3d registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10645–10662, 2024.

[114] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 21875–21911. Curran Associates, Inc., 2024.

[115] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[116] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024.

[117] Khaled Elgeneidy, Peter Lightbody, Simon Pearson, and Gerhard Neumann. Characterising 3d-printed soft fin ray robotic fingers with layer jamming capability for delicate grasping. In *2019 2nd IEEE International Conference on Soft Robotics (RoboSoft)*, pages 143–148, 2019.

[118] Krishna Shankar, Mark Tjersland, Jeremy Ma, Kevin Stone, and Max Bajracharya. A learned stereo depth system for robotic manipulation in homes. *IEEE Robotics and Automation Letters*, 7(2), 2022.

[119] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[120] Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[121] Inc. Jacobi Robotics. Jacobi motion library – next generation motion planning, 2024. https://docs.jacobirobotics.com.

[122] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[123] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, 2023.

[124] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for Gaussian splatting. *arXiv preprint arXiv:2409.06765*, 2024.

[125] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217, 2009.

[126] Jonathan Heewon Yang, Dorsa Sadigh, and Chelsea Finn. Polybot: Training one policy across robots while embracing variability. In *Conference on Robot Learning*, pages 2955–2974. PMLR, 2023.

[127] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. 2024.

[128] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

[129] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[130] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[131] Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspiar Singh, Clayton Tan, Jodilyn Peralta, Brian Ichter, et al. Scaling robot learning with semantically imagined experience. *Robotics: Science and Systems*, 2023.

[132] Zoey Chen, Sho Kiami, Abhishek Gupta, and Vikash Kumar. Genaug: Retargeting behaviors to unseen situations via generative augmentation. *Robotics: Science and Systems*, 2023.

[133] Nicholas Pfaff, Evelyn Fu, Jeremy Binagia, Phillip Isola, and Russ Tedrake. Scalable real2sim: Physics-aware asset generation via robotic pick-and-place setups. 2025.