Dynamic Object Understanding and Enhanced Grasp Detection: A Dual-Method Approach

Soofiyan Atar

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2024

Committee:

Joshua Smith

Linda Shapiro

Program Authorized to Offer Degree:

Electrical Engineering

University of Washington

Abstract

Dynamic Object Understanding and Enhanced Grasp Detection: A Dual-Method Approach

Soofiyan Atar

Chair of the Supervisory Committee:

Prof. Joshua Smith

Electrical Engineering

In this thesis, we present a comprehensive approach to advancing suction grasp point detection through several innovative methods. Initially, we introduced DYNAMO-GRASP, a novel technique leveraging the strengths of physics-based simulation and data-driven modeling to account for object dynamics during the grasping process. This method significantly enhances a robot's ability to handle previously unseen objects and scenarios in real-world settings, achieving a remarkable success rate improvement of up to 48% over state-of-the-art (SOTA) methods in challenging real-world tests. Building on this foundation, we elevated DYNAMO-GRASP by integrating Google-scanned objects with RGB channels, which further increased

accuracy by 30%. We also explored Visual Language Model (VLM) methods but found that they underperformed compared to the enhanced DYNAMO-GRASP RGB version, as they sometimes missed the suction grasp despite extensive prompt engineering efforts. Subsequently, we investigated zero-shot transfer using the ChatGPT VLM model.

The culmination of our research is the development of a hybrid model combining Dino V2 and DPT models. In this model, Dino V2 serves as the encoder and DPT as the decoder, with a complex head predicting the affordance map for grasp point extraction. This method has demonstrated the highest performance to date, doubling the accuracy of previous approaches. Additionally, it outputs roll and pitch affordance maps, which are used to determine the optimal grasping angles. This advanced model, validated using simulated data and transferred to real-world applications, marks a significant milestone in robust and resilient robotic manipulation in intricate real-world situations.

Acknowledgements

This thesis is the culmination of a journey that would not have been possible without the support and guidance of many individuals and organizations.

First and foremost, I would like to extend my deepest gratitude to Amazon Science HUB for their generous funding, which provided the essential resources and environment for my research.

I am deeply indebted to my advisor, Professor Joshua Smith, whose expertise, patience, and insightful feedback have been invaluable. His guidance and encouragement have been crucial throughout the course of this research.

I would also like to express my sincere thanks to Dr. Boling Yang, whose assistance in the initial stages of my research was instrumental in shaping the direction of my work. His knowledge and support were greatly appreciated.

Thanks to Prof. Maya Cakmak, Yi Li, Michael Murray, Nick Walker, and Dr. Markus Grotz for their collaboration and valuable contributions. Their discussions and insights helped to refine my ideas and improve the quality of this thesis.

Finally, I am profoundly grateful to my family and friends for their

unwavering support and encouragement. Their belief in me has been a constant source of strength and motivation.

Thank you all for your support and contributions to this journey.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Grasp point detection is essential for successful robotic manipulation, as it requires identifying the optimal location on an object for a robot to securely grasp and manipulate. Rapid and reliable grasping capabilities for a wide range of objects can benefit various applications, such as warehouse and service robots. Suction grasping is a popular grasping modality in real-world settings due to its simplicity and reliability when handling objects with non-porous, flat surfaces compared to parallel-jaw or multi-finger grasping. Existing methods for finding suitable grasping areas for suction grippers typically focus on maximizing suction seal quality and robustness against wrenches, taking into account the object's shape, size, and surface properties.

Most existing methods for suction grasping assume a top-down manipulation setting, where objects are initially placed on a stable, flat surface before being grasped, and the robot attempts to grasp the object from above. This is due to the suction cup gripper requiring the robot to apply a specific amount of force to press the suction cup against the object's surface, which causes

the cup to deform and create an air seal, resulting in a secure suction grasp. Consequently, an object being grasped needs sufficient and stable support in the direction opposite the robot's pushing. Without such support, the object may move in an unfavorable direction, leading to the suction cup's failure to form the air seal. However, numerous real-world scenarios require a robot to grasp objects without stable support, such as grasping from a container with a side opening or from an unstable pile of objects. In these situations, the objects may exhibit significantly more complex dynamics during the manipulation process due to the displacement caused by the robot's motion and the objects' interactions with one another. State-of-the-art grasp point detection methods for suction grasping could suffer from these complex object-picking scenarios because they do not consider the objects' movement during the manipulation process. This limitation greatly restricts the range of scenarios in which suction grippers can be applied, preventing them from reaching their full potential in real-world manipulation tasks.

To address these challenges, we developed DYNAMO-GRASP [1], a novel technique leveraging the strengths of physics-based simulation and data-driven modeling to account for object dynamics during the grasping process. This method significantly enhances a robot's ability to handle previously unseen objects and scenarios in real-world settings, achieving a remarkable success rate improvement of up to 48% over state-of-the-art (SOTA) methods in challenging real-world tests.

Building on this foundation, we elevated DYNAMO-GRASP [1] by integrating Google-scanned objects with RGB channels, which further increased accuracy by 30%. We also explored Visual Language Model (VLM) methods but found that they underperformed compared to the enhanced DYNAMO-GRASP RGB version, as they sometimes missed the suction grasp despite extensive prompt engineering efforts. Subsequently, we investigated zero-shot transfer using the ChatGPT VLM model.

The culmination of our research is the development of a hybrid model combining Dino V2 and DPT models. In this model, Dino V2 serves as the encoder and DPT as the decoder, with a complex head predicting the affordance map for grasp point extraction. This method has demonstrated the highest performance to date, doubling the accuracy of previous approaches. Additionally, it outputs roll and pitch affordance maps, which are used to determine the optimal grasping angles. This advanced model, validated using simulated data and transferred to real-world applications, marks a significant milestone in robust and resilient robotic manipulation in intricate real-world situations.

This thesis makes the following contributions:

1. **Suction Grasping by Taking Object Movement into Consideration:** We describe the challenge of complex object movement during suction grasping, which no current state-of-the-art method adequately addresses.

2. **An Open Source Novel Suction Grasping Simulation:** To address this challenge, we developed a high-performance suction grasping simulation environment using Isaac Gym. This simulation environment models the influence of object dynamics on the success of suction grasps throughout the grasping process.

3. **A Dataset and Learned Model:** Utilizing the simulation environment, we generated a dataset that contains more than one million simulated grasps and trained a grasp point detection model that takes into account how the movement of objects and the robot's kinematics impact the success of grasping.

4. **Evaluation in a Real-world Warehouse Setting:** We assessed two grasp point detection approaches alongside our model. In both simulated and real-world experiments, our method surpassed the alternatives in terms of accuracy and consistency.

The results of this research set the stage for more reliable and resilient robotic manipulation in intricate real-world situations.

Chapter 2

Related Work

Suction-based robot manipulators have gained widespread popularity in real-world applications. For instance, suction grasping methods are used in manufacturing [2, 3, 4], warehousing [5, 6], underwater manipulation [7, 8], food and fruit manipulation [9, 10, 11, 12], etc. Another major direction where suction grasping has been applied is in the exploration of end-effector modalities [13, 14, 15, 16].

2.1 Suction Grasping Techniques

Analytic Models: In the realm of conventional suction cup grippers, the effective analysis of grasp quality necessitates the modeling of various cup properties. Given that these suction cups are typically fashioned from elastic materials, such as rubber or silicone, researchers frequently employ spring-mass systems to represent their deformations [17, 18, 19]. Upon establishing a secure grasp on an object using a suction gripper, the suction cup is typically modeled as a rigid entity. The subsequent analysis involves assessing the

forces imposed on the object, encompassing those along the surface normal, friction-induced tangential forces, and suction-generated pulling forces [20]. Mahler et al. [17] introduced a combined model in DexNet3.0, incorporating both torsional friction and contact moment within a compliant model of the contact ring between the cup and the object. This amalgamated model has demonstrated its efficacy and is employed in subsequent works [18, 21]. Additionally, this work adapts the analytic models from DexNet3.0 for the purpose of data annotation.

Learning Suction Grasps: Machine learning research in robotics has been actively exploring the selection of optimal grasp points to enhance suction grasping for intricate manipulation tasks [22, 23]. These tasks include novel object picking, object stewing, picking from containers, etc. Existing approaches generate training data through either human expertise [24] or simulations [17, 18, 22, 25]. DexNet3.0 [17], for instance, synthesizes training data and proposes suction grasp points that aid in forming an effective suction seal and ensuring wrench resistance. Several other studies center around clustered scenarios by creating models that take RGB-D input and predict graspable points [18, 25, 24]. Jiang et al. [22] proposed a methodology that simultaneously considers grasping quality and robot reachability for bin-picking tasks. Despite these studies primarily focusing on analyzing surface properties or robot configuration, they largely overlook how the displacement of the object during the picking process might impact the success of the task. Addressing

this particular aspect is the main focus of our work.

Visual Pushing: This project also relates to the active research area of object displacement modeling duringThis project also shares relevance with the active research area of object displacement modeling during manipulation [26, 27]. Effective non-prehensile manipulation strategies have been successfully applied to enhance grasping operations [28, 29, 30]. Recently, reasoning object translation via visual input has gained huge advances. Transporter and its variants [31, 32] have introduced a data-efficient learning paradigm that links visual inputs to desired robotic actions. Nonetheless, these methods are underpinned by a strong assumption of translational equivariance in visual representation, a condition that is often not met in non-table-top settings. Visual foresight methods [33, 34] have offered a model-based framework that predicts future observations based on a state-action pair. However, these approaches necessitate searching through the action space given a specific task, which can be time-consuming for intricate real-world problems. Other existing studies [35, 36, 37] have examined robotic manipulation from a sideways perspective. However, none of them have explicitly modeled the complex dynamics caused by the interaction between the robot and the objects.

2.2 VLM-Based Grasp Point Extraction

Visual Language Models (VLMs): Visual Language Models have emerged as powerful tools for understanding and generating language-augmented vi-

sual content. Recent studies have shown that VLMs can be leveraged to enhance robotic grasping tasks by using natural language instructions to guide the grasp point detection process [38, 39]. These models combine visual input with contextual language cues to identify graspable points on objects. While promising, VLM-based methods often face challenges related to the accuracy and reliability of grasp point detection in dynamic and cluttered environments. Additionally, prompt engineering is critical to improving the performance of VLMs in specific grasping scenarios, but it can be labor-intensive and sometimes ineffective.

Zero-Shot Learning for Grasping: Zero-shot learning techniques aim to generalize grasping models to novel objects without requiring extensive retraining. Recent advancements have employed large-scale visual language models, such as ChatGPT VLM, to perform a zero-shot transfer of grasping knowledge [40]. These approaches enable robots to understand and execute grasping tasks based on high-level descriptions and visual cues, significantly reducing the need for extensive labeled data. However, the performance of zero-shot VLM-based grasping methods can vary, especially when dealing with objects that exhibit complex geometries and dynamic behaviors.

Chapter 3

DYNAMO GRASP

3.1 Introduction

Traditional methods for suction grasping typically focus on maximizing the quality of the suction seal and ensuring robustness against various forces, considering the object's shape, size, and surface properties.

However, most existing suction grasping techniques are limited by their assumption of a top-down manipulation approach. In this scenario, objects are placed on stable, flat surfaces before being grasped, allowing the robot to apply the necessary Force to deform the suction cup and create an air seal. This method requires stable support for the object in the direction opposite to the robot's pushing Force. Without such support, the object may shift, causing the suction seal to fail. In many real-world situations, robots must grasp objects without stable support, such as from a container with a side opening or an unstable pile of items. These scenarios introduce complex dynamics during manipulation due to object displacement and interactions.

Current state-of-the-art methods for suction grasping often fail in these complex scenarios because they do not consider the movement of objects during the grasping process. This limitation significantly restricts the applicability of suction grippers in diverse real-world tasks, preventing them from achieving their full potential.

To overcome these challenges, we developed DYNAMO-GRASP, a novel approach that integrates physics-based simulation and data-driven modeling to account for object dynamics during grasping. By considering how objects move and interact, DYNAMO-GRASP enhances the robot's ability to handle previously unseen objects and scenarios, substantially improving success rates over traditional methods.

This chapter outlines the development and evaluation of DYNAMO-GRASP, focusing on the challenge of complex object movement during suction grasping, which is inadequately handled by current methods.

3.2 Problem Statement

Our objective is to identify grasp points on a target object within a container filled with multiple items using a single-view depth image observation. The identified grasp points should enable a robot to successfully establish a suction grasp, even when the object lacks stable support in the direction opposite to the robot's push. Consistent with previous suction grasp point detection studies, a grasp point is defined by a target point $[p, v]$. Here, $p \in \mathbb{R}^3$ repre-

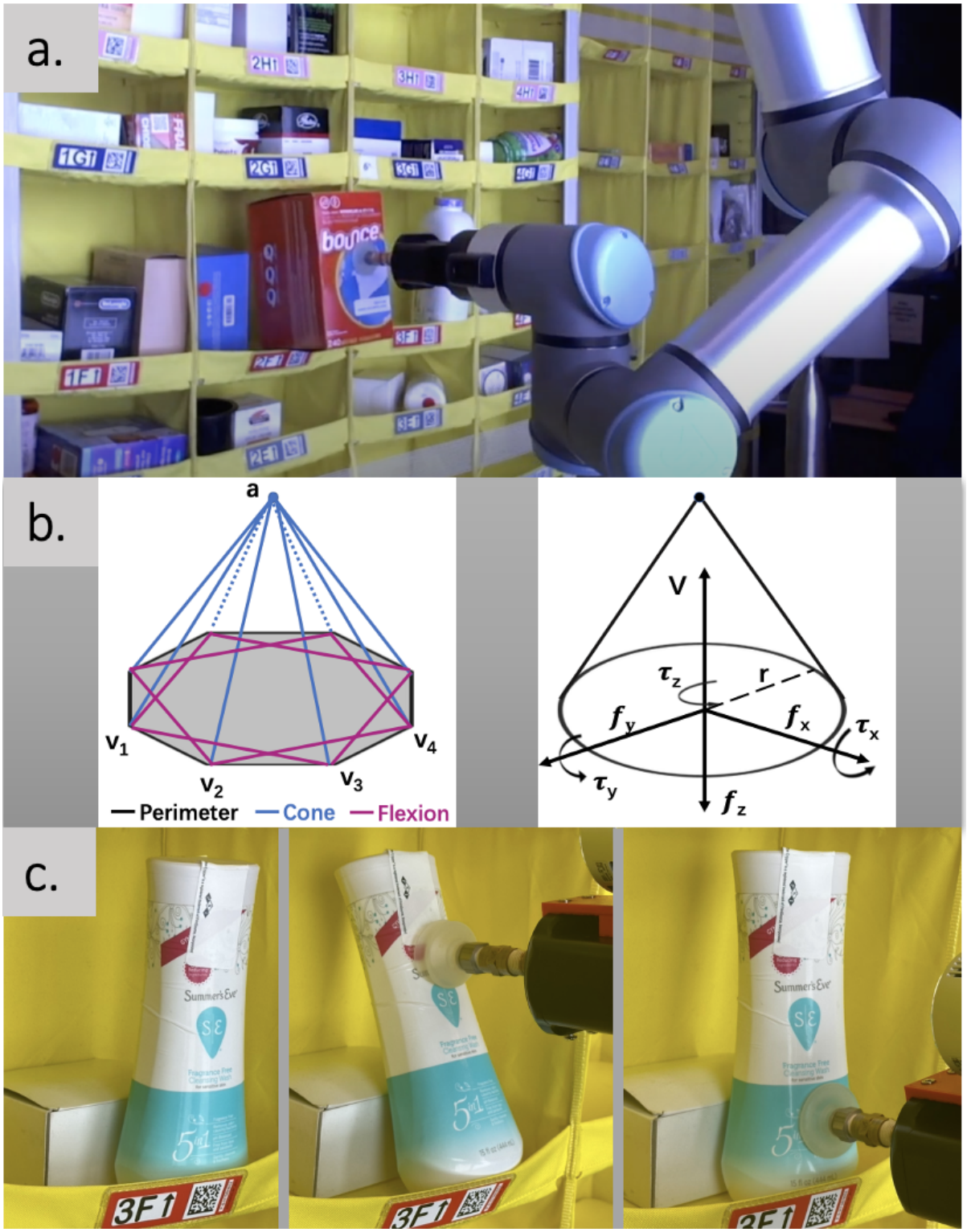

Figure 3.1: a. Suction grasping for real-world scenarios remains challenging due to limited analysis of object movements. b. SOTA methods only reason for object's surface properties. *Left*: The quasi-static spring model. *Right*: Wrench basis for the suction cup. [17] c. *Left*: A warehouse picking scenario. *Middle*: DexNet failing the grasp due to object toppling. *Right*: An effective grasp point that prevents unfavorable object movements. See the project website for experiment videos.

sents the center of the contact ring between the suction cup and the object, while $v \in S^2$ denotes the gripper's approach direction. The grasp labeling function is defined as 1 if the grasp successfully forms a suction grasp on the target object and 0 if it does not. This section discusses the crucial criteria for forming a successful suction grasp.

3.2.1 Seal Quality and Wrench Resistance

A suction cup is capable of lifting objects owing to a differential in air pressure. This differential is created across the cup's membrane by a vacuum generator, which pulls the object toward the cup. Ensuring a tight seal between the suction cup and the target object is crucial for successful operation. For sealing evaluation, we follow the highly effective quasi-static spring-based model proposed in DexNet 3.0. This model uses a combination of three spring systems to represent suction cup deformation. A perimeter spring system assesses the deformation between adjacent vertices, namely v_i and v_{i+1}. The cone spring system signifies the deformation of the suction cup's physical structure, as determined by the distance between v_i and a. Lastly, the flexion springs, which connect vertex v_i to v_{i+2}, are employed to resist bending along the surface of the cup.

When the suction cup forms an air seal with the object, the suction gripper should be able to resist wrenches caused by gravity or other disturbances. The suction ring contact model proposed in DexNet 3.0 efficiently encapsulates

the forces experienced by a suction cup during a grasp. This model considers five forces: the actuated normal Force (f_z) and vacuum force (V) represent the gripper pressing into the object along the contact axis and air suction pulling the object, respectively. The friction forces (f_x, f_y) and torsional friction (τ_z) result from the normal Force exerted between the suction cup and the object, acting as resistive forces. Lastly, the elastic restoring torques (τ_x, τ_y) result from the elastic restoring forces within the suction cup, which apply torque on the object along the boundary of the contact ring.

3.3 Object Movement

Most established suction grasping techniques assume little to no movement of the object during the process, which facilitates the deformation of the suction cup, thereby enabling the formation of an air seal for a secure grasp. However, in various practical manipulation scenarios, the target object might not have ample and steady support opposite the robot's push. This lack of support can lead to undesirable shifts in the object's position, preventing the successful creation of the air seal. The situation becomes even more complex when other objects are located near the target due to the interactions among them. This work addresses these complexities by modeling the movement of objects during the picking process, which enhances the applicability and efficiency of suction-based grippers in real-world manipulation tasks. Assuming that an object's state is denoted by its Cartesian pose and velocity in a workspace,

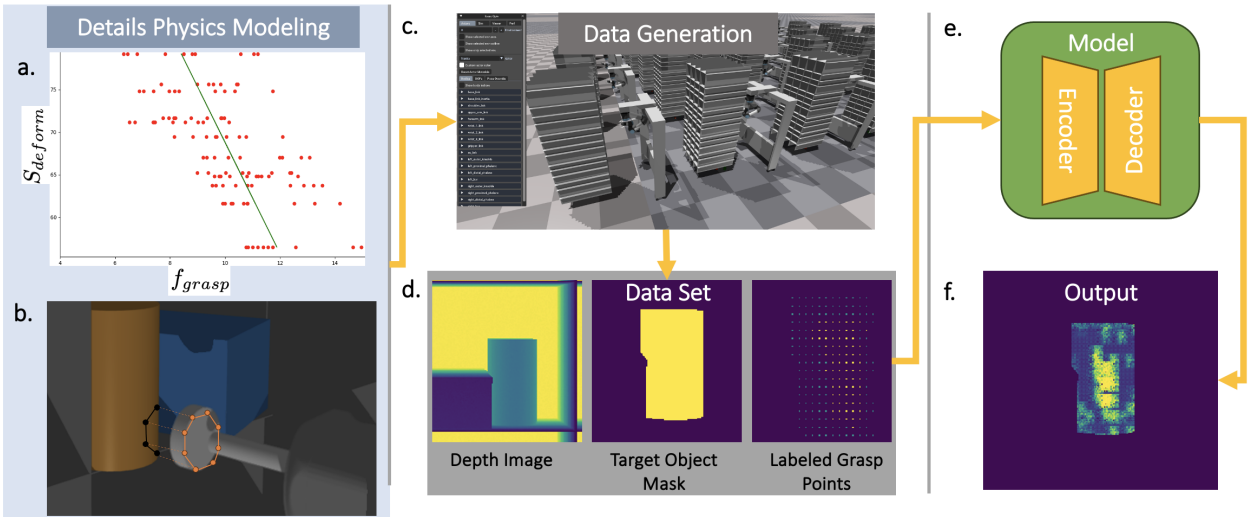

Figure 3.2: An overview of the proposed pipeline: **a.** We conducted system identification using 19 everyday objects of diverse shapes, weights, volumes, and materials to ascertain the function F discussed in Section 3.4. **b.** Calculation of deformation score at each simulation time step. **c.&d.** Generating dataset with our simulation environment. **e.&f.** Trained DYNAMO-GRASP model outputs an affordance map highlighting optimal grasp areas.

represented as $s = (p, \delta p)$, the states of i objects in a container at time t can be represented as $s_t = \{s_{t0}, s_{t1}, ..., s_{ti}\}$. At each time step, a robot equipped with a suction gripper performs a pushing action $a_t = (f_t, p, v)$, applying a force f_t to a specific location, p, on the object's surface in the direction of v. The state transition model $p(s_{t+1}) = T(s_t, a_t)$ provides a distribution over the potential movements of the objects during the picking process.

DYNAMO-GRASP aims to improve the success rates of suction grasps in complex, real-world scenarios by addressing the challenges related to seal quality, wrench resistance, and object movement.

3.4 DYNAMO-GRASP

This section proposes a robot learning pipeline designed to create a grasp point detection model. This model suggests suction grasp points by analyzing combined information regarding object surface properties and object movement during the picking process. We first implemented a new suction grasping simulation environment that accurately simulated suction cup properties and objects' displacement caused by the robot's motion and the objects' interactions with one another. A transformer-based model is trained to take a depth image as input and generates an affordance map over the target object's surface, indicating the likelihood of a successful suction grasp if a robot executes a pushing action along the surface normal across various areas of the object. Please note that our method primarily focuses on analyzing the impact of physical interaction between robots and objects on the quality of a suction grasp. During execution, we filter out grasp points that offer inadequate air seals and wrench resistance based on DexNet's output. Fig.3.2 shows the system architecture.

3.5 Simulation Environment and Data Generation

In order to eliminate the need for expensive real robot data collection, we carefully designed a simulation environment that accurately replicates the physical properties of the suction cup, the motion of objects caused by robot

grasping, as well as the robot's kinematics during the picking process. We chose to implement our grasping simulation environment based on Isaac Gym, allowing all computations to be accelerated via GPUs. While Isaac Gym lacks important features that emulate detailed suction grasping properties, our environment integrates several custom-implemented functional modules. It provides a pipeline that accurately simulates a suction-picking process by taking into account factors such as suction cup properties, robot kinematic constraints, collisions, control noise, and object dynamics.

Modeling Suction Properties. The majority of popular physics simulations for robotics merely simulate suction grasp through simplistic mechanisms. These mechanisms typically involve directly attaching the object to the robot's end-effector or creating an attracting force between the object and the effector. However, these approaches neglect critical physical details. Specifically, to successfully register a suction grasp, the suction cup must be pushed and deformed to a sufficient extent that the rim of the suction cup attaches to the surface of the object, thereby forming an air seal. Modeling the amount of Force required to form an air seal is crucial for this problem. This is because when the target object lacks rigid support, exerting sufficient Force directly causes the object's displacement. Understanding the magnitude of the Force that the robot exerts on the target object is instrumental in recreating accurate object dynamics. To model the deformation properties of the suction cup, we first adopted the Perimeter Springs in the

quasi-static spring system. Given a grasp point $\mathbf{p}$ on the object's surface and the angle of incident $\mathbf{v}$, this model calculates a suction deformation score $S_{deform} = 1 - \max(r_1, r_2, ..., r_n)$, where $r_i = \min(1, |(l_i' - l_i)/l_i|)$. Here, l_i represents the original length of the perimeter spring linking vertex v_i and v_{i+1}, and l_i' is the length after projecting the vertices onto the object's surface. Using real-world data, we then conduct a system identification process to ascertain the function F. This function signifies how forcefully the robot needs to press the suction cup to achieve a successful grasp, given a deformation score of a specific grasp point: $F(S_{deform}) \rightarrow f_{grasp}$.

Simulating Grasping Physics. *(1) Kinematics:* Our simulation accepts a robot's model as input and controls the robot using an end-effector controller to attempt various suction grasps. This approach enables the simulation to demonstrate how the robot's form factor and kinematic properties impact its grasp. For instance, some grasp points might be physically unattainable for the robot due to its manipulability and reachability constraints or collisions. *(2) Generating Scenarios:* Our experiment primarily focuses on a warehouse lateral picking scenario. During our data generation process, the simulation randomly selects one to three objects from our object set and spawns them into the same container with random positions and orientations. We also implement domain randomization for observation noise, objects' weights, and controller parameters, ensuring the dataset reflects a range of diverse physical properties and robot behaviors. One of the objects in the container is ran-

domly assigned as the target object to be picked. *(3) Sampling Grasp Points:* Given a picking scenario, we sample two sets of candidate grasp points from the visible surface of the target object. The first set is derived from uniform sampling across the entire surface, ensuring that the robot explores diverse picking strategies. The second set contains the grasp points with the highest score returned by DexNet via the Cross-Entropy Method (CEM) sampling strategy, ensuring the robot explores areas that DexNet deems preferable.

Labeling Data. After sampling the candidate grasp points, our simulated robot 'physically' executes each candidate $\mathbf{p}$ by performing a sequence of pushing actions $\mathbf{A} = \{a_t\}_{t=0}^{T}$, where each action $a_t = (f_t, \mathbf{p}, \mathbf{v})$. Here, $\mathbf{v}$ is determined by the surface normal at $\mathbf{p}$. The robot exerts a constant force $f_t = f_c$ if the target object is unstable and moves in response to the gripper's push. Once the object finds a position with adequate support against the push, f_t gradually increases until the suction cup deforms enough to form an air seal or until the object starts moving again. The simulation of the suction cup's deformation and the precise estimation of suction grasp registration involves a continuous calculation of S_{deform} and f_{grasp} at each timestep. This process considers the suction cup's current position relative to the target object, as shown in Fig.3.2.b. A force sensor on the end-effector continuously monitors f_t, and a grasp is deemed successful if $f_t \geq f_{grasp}$. Any failure to meet this condition, such as collisions or inaccuracies in end-effector positioning due to manipulability or reachability issues, results in the grasp point being marked

as unsuccessful. For successful grasps, we incorporated a penalization term p_{move} into the label to penalize unnecessary object movements.

3.5.1 Model Training

We employ the suction grasping simulation, as described above, to generate a dataset. This dataset represents a warehouse scenario where a robot equipped with a suction gripper is tasked with extracting a target object from a small container filled with multiple unorganized items. The experimental setting is detailed in Sec.3.9. This dataset was used to train a model for grasp point selection. The model takes a single-view point cloud of the container's interior, a segmentation mask identifying each object within the container and its boundaries as inputs. It then outputs an affordance map representing the estimated probability of successful grasps at all potential grasp points on the target object. The largest value in the affordance map indicates the optimal grasp point, $(\mathbf{p}^*, \mathbf{v}^*)$, for the given scenario. As shown in Fig.3.2.e, our model employs an auto-encoder architecture, integrating a transformer encoder and a deconvolutional decoder. As previously mentioned, our data generation process is designed to capture the inherent variabilities of complex real-world robotic suction grasping tasks. These include variations stemming from the physical properties of different objects, robot constraints, and stochasticity in the controller, among others. Empirically, we discovered that the following loss function can effectively mitigate the adverse effects of the high aleatoric

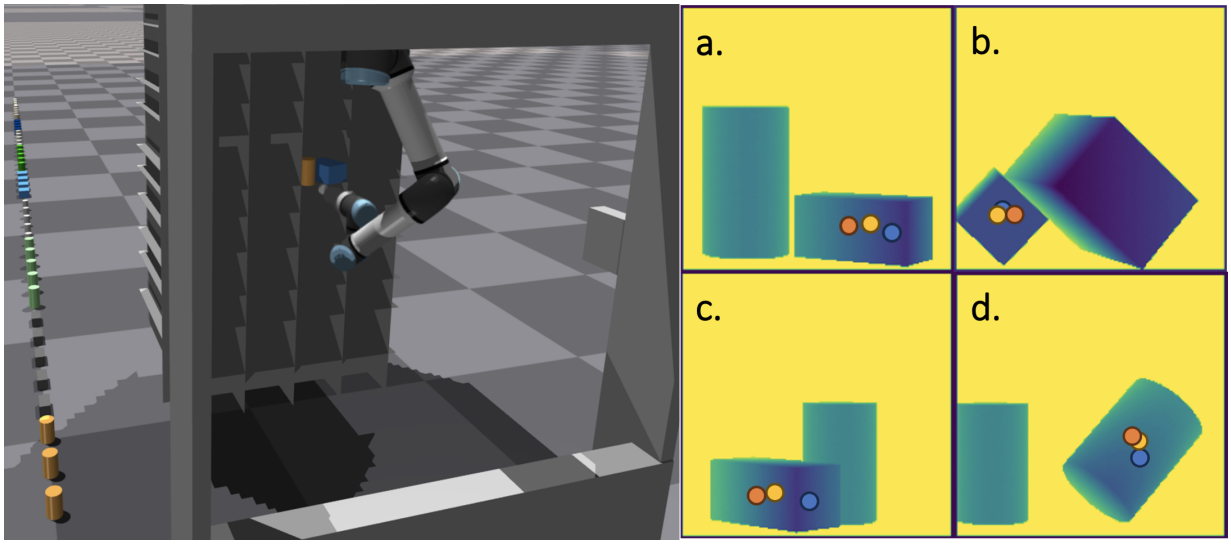

Figure 3.3: **Left:** The simulation environment for data generation and experiments. The simulated objects with different weights, sizes, and shapes are displayed on the left side of the robot. **Right:** In Section 3.9.1, challenging test cases are presented where only DYNAMO-GRASP was successful in grasping the target object. The orange, blue, and yellow points indicate the grasp points proposed by DYNAMO-GRASP, DexNet, and the Centroid method, respectively.

uncertainty in our dataset during training: $\mathcal{L}_{Y_{max}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \forall y_i \in Y_{max}$. Here, Y_{max} represents a subset of samples containing the n highest-scored grasp points on an object.

3.6 Simulation Details

A simulation that accurately replicates the targeted robotic tasks can significantly enhance the efficiency of various machine learning algorithms in learning these tasks[41, 42, 43, 44]. Most physical simulators, such as Mujoco[45], PyBullet[46], and IsaacGym[47], which excel at simulating the physical properties of object motion, lack the functionality to simulate the characteristics of suction cups during suction grasping. Our development efforts focus on

utilizing the sensing and physical information in IsaacGym to create more realistic suction-picking properties.

System Identification: Our system identification process aims to accurately model the Force required by the robot to deform the suction cup. This ensures the rim of the cup adheres to the object's surface, forming an air seal. We chose 18 everyday objects with varied surface geometric characteristics, aiming to cover a broad spectrum of deformation scores. For each object, we executed ten suction grasps using our UR16 robot. To minimize measurement noise, the objects were held firmly to limit movement during the grasping process. The Force required for the suction gripper to achieve a suction seal was detected by a sudden decrease in suction airflow and the force torque sensor located on the robot's wrist. We observed that the characteristics of our suction cup differ significantly between nearly flat object surfaces and those that are more curved or intricate. Consequently, we chose to represent the function $F(S_{deform}) \rightarrow f_{grasp}$ using a hybrid linear function:

$$F(S_{deform}) = \begin{cases} 7.66 - 0.06 * S_{deform} & \text{if } S_{deform} \leq 80 \\ 22.2 - 0.18 * S_{deform} & \text{otherwise} \end{cases}$$

Typically, the size and firmness of a suction cup influence its working range for objects of varying sizes and weights. However, this doesn't profoundly alter the nature of this grasping problem. For instance, when using a small suction cup to manipulate lighter, smaller objects, these objects typically

have less friction with the container and reduced inertia, making them more prone to toppling. However, even though we anticipate a certain degree of generalization to unseen suction cups, we recommend carrying out the system identification process to achieve optimal performance.

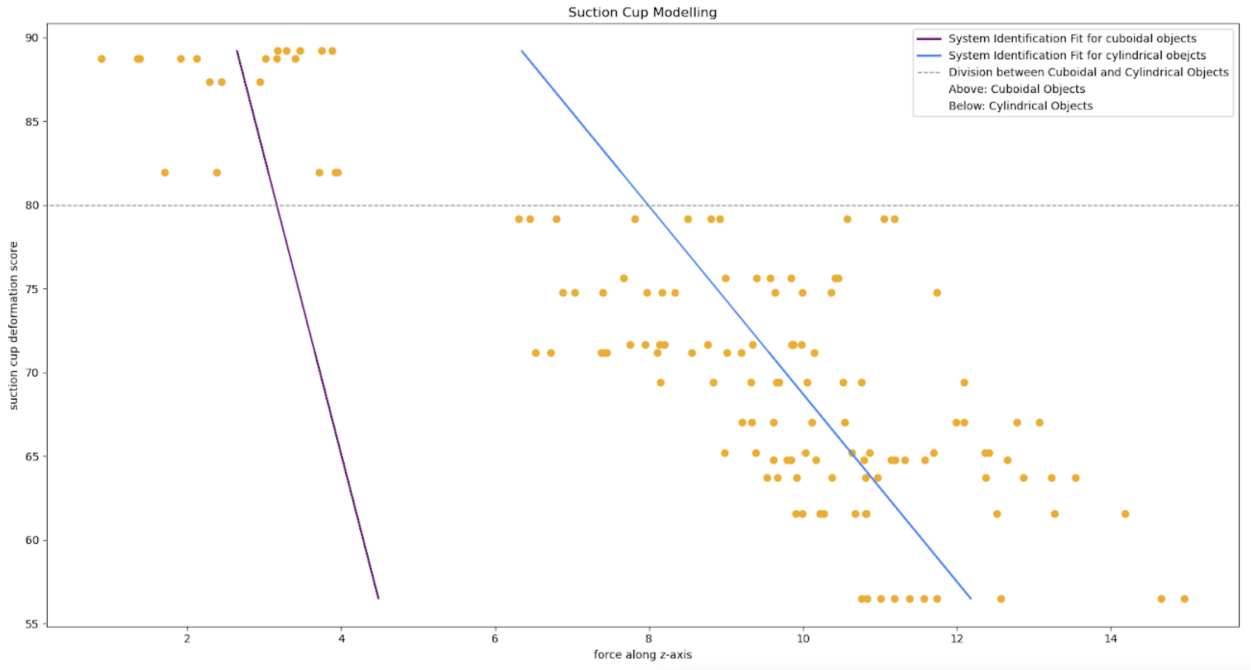

Figure 3.4: Force exerted on an object as a function of the suction deformation score. Solid lines represent system identification fits for cylindrical (blue-colored line) and cuboidal (violet-colored line) objects. The dotted line demarcates the distribution of data points between the two object types.

Domain Randomization: We performed domain randomization to vary object weights, where each of the ten chosen objects had their original weights varied by -5, -10, +5, and +10, leading to five weight versions per object. These objects were cylindrical or cuboidal with varying dimensions, inertia, and weights. The cylindrical ones had radii from 26 to 46.3, heights between 95 and 155, and weights ranging from 118 to 602. In contrast, the cuboidal objects had lengths from 112 to 165, breadths between 55 and 102,

and widths from 55 to 95. Notably, with every weight change, the inertia properties were appropriately modified. When placing objects in the simulator, their orientations on all three axes were uniformly picked from -180 to 180. Although their initial placements followed predefined bin coordinates, potential collisions might displace some objects. As a preventive measure, we ensured that each object remained within the bin limits and verified the stability of each object setup by spawning it thrice and monitoring its movement at each simulation step for minimal displacement until the suction gripper came in contact with the target object. Lastly, to closely mimic our physical robot setup with the Intel RealSense L515 camera, we added Gaussian noise (mean: 0, standard deviation: 0.9) to the depth images.

Labeling: For the label for each configuration, each grasp point score serves as an indicator of grasp success. A grasp point that fails to achieve a secure suction grip is assigned a definitive zero score. Additionally, the label is designated as a 'failure' if the robotic arm does not align and picks the object at the computed angle of incidence derived from the surface normals, ensuring the grasp adheres to the pre-calculated optimal orientation. Another critical constraint is that the arm must avoid unintended contact with any other object before establishing contact with the target, as such collisions can compromise the grasp's integrity and lead to potential inaccuracies or damage. On the other hand, successful grasp points are scored using the

equation $s = 1 - p_{move}$, where,

$$p_{move} = \max(0, \min(obj\_movement, 0.3))$$

$$obj\_movement = \sum_{t=0}^{T-1} (||\text{tran}_{t+1} - \text{trans}_t|| + (1 - |\text{quat}_{t+1} \cdot \text{quat}_t|))$$

p_{move} is a penalization term that discourages unnecessary movement of the target object. *obj_movement* calculates the total movement of the target object during the picking process. The picking horizon T is discretized by a fixed interval, and t represents a time step within T. tran and quat represent the translation and orientation of the target object at a given time step, respectively.

Dataset: We implemented specific data augmentation techniques on our dataset to enhance our model's resilience against variances in real-world scenarios. We added Gaussian noise to the point cloud data and flipped the input data along with their corresponding labels, strengthening the model's ability to recognize various object orientations and thereby improving its generalization capabilities. These augmentation strategies significantly expanded the diversity of our training dataset, ensuring the model's proficiency in managing diverse input perturbations. The complete dataset, including labels and augmented inputs, consisted of around 12000 configurations, including augmentations, enhancing the dataset's diversity and depth, which occupy approximately 10 GB of storage space.

3.7 Learning Details

Model Architecture: Our model employs a variation of the Vision Transformer (ViT), adopting the architecture from Beyer et al.[48]. We chose ViT because it represents a state-of-the-art architecture widely used in vision classification tasks. Utilizing this architecture demonstrates that a standard network, when trained with our dataset, effectively addresses the challenge of suction grasping in complex object clusters. This is achieved without the need for custom modifications to the model architecture.

Loss Fuction: We initially experimented with both the standard MSE loss and Cross-entropy loss but observed only mediocre performance from the model. As highlighted in Section 4, the domain randomization process introduced significant stochasticity to our dataset. Empirically, we found the presented loss function to be more effective in this specific context. The use of Y_{max} is a simple technique designed to mitigate the adverse effects of high aleatoric uncertainty present in the training data. It updates the model by only taking into account the grasp points that the model deems to have a high confidence of success. This approach is aimed at penalizing false positive predictions made with high confidence or encouraging true positive predictions made with high confidence while disregarding low confidence labels, which usually arise due to data noise. We discovered that this loss function led to improved prediction accuracy and produced a smoother affordance map.

Hyperparameters: We trained our ViT model using the Adam optimizer with a learning rate of $5e^{-5}$ and a batch size of 128 images. The model converged in about 500 epochs, and the training was conducted on an NVIDIA RTX 3090. Our ViT model consists of eight heads, each with a dimension of 64. Consequently, we set Q, K, and V to 128, 257, and 1536, respectively. The model accepts a $4 \times 256 \times 256$ tensor as input. The first, second, and third channels represent the x, y, and z values of the cropped point cloud observation for the container. The fourth channel provides a segmentation mask that localizes the target object. The model produces a 256×256 affordance map. Each pixel in this map provides a score ranging from 0 to 1. A higher score indicates a more favorable grasp point for achieving a successful suction grasp.

Segmentation Mask: Within the Isaac GYM simulator, we adhere to ground truth segmentation masks. For real-robot experiments, we used a specialized method, called "STOW" [49], that combines VITA [50] and the Mask2Former [51], which is tailored for joint unseen object instance segmentation and tracking. The method uses transformer-based architectures and dynamic tracking anchors to handle real-world visuals characterized by dense clustering and substantial intra-frame object displacements.

Grasp Point Selection: After getting the affordance map, we first identify the pixels that represent the target object in the map using the segmentation mask. Subsequently, we use the DBSCAN algorithm [52] to cluster regions

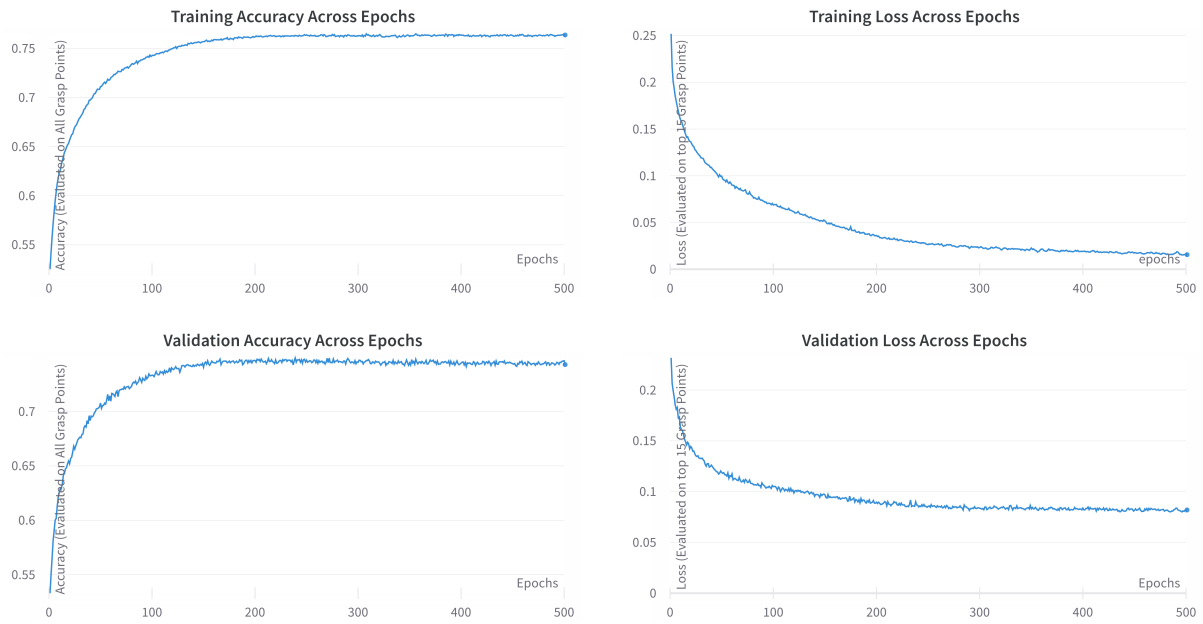

Figure 3.5: Training and validation metrics over epochs: The top row displays the metrics related to training, with the left graph showing the training accuracy (calculated using all grasp points) and the right graph presenting the training loss (determined with 15 highest-scored grasp points). The bottom row focuses on validation metrics, with the left graph illustrating the validation accuracy (using all grasp points) and the right graph depicting the validation loss (using the 15 highest-scored grasp points)

displaying high-affinity scores exceeding 0.9. During this clustering phase, each cluster encompasses a minimum of five pixels. The final grasp point is defined by the centroid of the cluster with the highest average affinity score.

3.8 Experiment Details:

In the real-world environment setup, distinct from the simulator approach, objects were first stowed into a designated bin. Following this, a segmentation algorithm was employed to generate a mask delineating each object. The user then selects the target object based on its unique value in the grayscale mask image, referred to as the 'target object id'. With the object identified,

the next phase involves running the inference of a user-provided algorithm to determine the optimal strategy for picking the selected object. The entire operation is orchestrated through a state machine, ensuring a seamless transition between stages. Each state is connected sequentially. In evaluating success and failure across various methods, a grasp point is deemed unsuccessful if motion planning fails consecutively on two occasions. Additionally, if the system does not create a suction with the object, it is also considered a failure. A successful grasp is solely determined by the creation of a good suction with the target object.

3.8.1 Extra Experimental Result:

The Sim2Real Gap: Our DYNAMO-GRASP model was exclusively trained using simulated data. In most of our experiments, this model exhibited outstanding real-world performance without requiring tuning using real-world data. This indicates the model's strong ability to generalize in real-world conditions, showcasing a minimal sim2real gap. To delve deeper into our pipeline's constraints, we pinpointed situations where simulation deemed the target object "impossible" to pick up. In these instances, all three picking techniques registered a 0% success rate in simulation. Importantly, these situations are infrequent, accounting for just around 3% of the 260 simulated test scenarios. We then mirrored these situations in an actual warehouse environment and ran a real robot experiment as delineated in Sec.3.9.2. Despite

Figure 3.6: Real-world adversarial evaluation with five grasp points for each configuration: DYNAMO GRASP (our method), DexNet, and Centroid. The color-coded points represent the suggested grasp points success and failure from various algorithms. The successfully identified grasp points are marked by the color along the label "success" and "failure".

the struggles faced by all three methods to secure high success rates (DYN: 16%, Dex: 8%, Cen: 40%), the real-world challenges weren't as formidable as projected by the simulation. However, this did highlight a sim2real gap in these rare scenarios. Our observations also revealed that, in cases where our simulation wasn't entirely accurate, the centroid method surpassed the performance of learning-based approaches. This observation emphasizes the value of refining learning-based models using actual world data.

Common Set						
	Real world experiments			sim experiments		
	DYN	Dex	Cen	DYN	Dex	Cen
Scenario 1	3	2	3	5	5	5
Scenario 2	5	4	5	5	5	5
Scenario 3	5	4	5	5	5	5
Scenario 4	5	5	5	5	5	5
Scenario 5	5	5	5	5	5	5
Scenario 6	5	5	5	5	5	5
Scenario 7	4	4	4	5	5	5
Scenario 8	5	4	4	0	5	5
Scenario 9	5	4	2	5	0	5
Scenario 10	5	5	4	5	5	5
Avg. Success Grasps	4.7	4.2	4.2	4.5	4.5	5
Std. Dev.	0.675	0.919	1.033	1.581	1.581	0
Total Success Rate	94%	84%	84%	90%	90%	100%

Table 3.1: Comparative evaluation of grasp success rates in common scenarios for three methodologies: DYNAMO-GRASP (DYN), DexNet (Dex), and Centroid (Cen). The table enumerates the average success rates, standard deviations, and total success rates for each method.

3.9 Experiment

Our experiment focuses on robotic suction grasping for industrial warehouse shelves [53]. Fig.3.1 depicts the robot setup and the industrial shelving unit which is packed with objects. The opening of these shelving units is located on the side, which makes suction grasping significantly more challenging compared to top-down manipulation scenarios, as the robot's movements can trigger a series of object displacements, leading to objects being shifted or even toppled over. Consequently, this scenario serves as an excellent evaluation environment for our work. Our system setup is as follows. Throughout our evaluation, we employed a Universal Robots UR16e robot equipped with a

Challenging Set						
	Real world experiments			sim experiments		
	DYN	Dex	Cen	DYN	Dex	Cen
Scenario 1	4	1	2	5	0	0
Scenario 2	2	2	3	0	3	3
Scenario 3	4	1	0	5	0	0
Scenario 4	0	1	3	0	0	2
Scenario 5	5	1	1	5	0	0
Avg. Success Grasps	3	1.2	1.8	3	0.6	1
Std. Dev.	2	0.447	1.304	2.739	1.342	1.414
Total Success Rate	60%	24%	36%	60%	12%	20%

Table 3.2: Comparative evaluation of grasp success rates in challenging scenarios for three methodologies: DYNAMO-GRASP (DYN), DexNet (Dex), and Centroid (Cen). The table enumerates the average success rates, standard deviations, and total success rates for each method.

Robotiq EPick suction gripper and an Intel Realsense L515 camera mounted on its wrist.

In our experiment, we focus on evaluating three methods:

1. our method DYNAMO-GRASP (Dyn),
2. DexNet3.0 (Dex), and
3. the Centroid method (Cen).

DexNet3.0 is a SOTA suction-picking technique, serving as a strong baseline. Meanwhile, the Centroid method, a straightforward approach involving suctioning on the object's centroid, has proven effective in similar tasks at the Amazon Robotics Challenge [37, 54].

Adversarial Set			
	Real world experiments		
	DYN	Dex	Cen
Scenario 1	5	3	2
Scenario 2	3	0	0
Scenario 3	1	0	1
Scenario 4	5	3	1
Scenario 5	5	1	3
Avg. Success Grasps	3.8	1.4	1.4
Std. Dev.	1.789	1.517	1.14
Total Success Rate	76%	28%	28%

Table 3.3: Comparative evaluation of grasp success rates in adversarial scenarios for three methodologies: DYNAMO-GRASP (DYN), DexNet (Dex), and Centroid (Cen). The table enumerates the average success rates, standard deviations, and total success rates for each method.

	Dyn(Full)	Dyn w/ MSR	Dyn w/o PEN	Dex	Cen
Total Success Rate	88.05%	86.75%	82.93%	81.12%	78.78%
Success Std	0.30	0.32	0.36	0.36	0.40

Table 3.4: The first row of the table displays the grasping success rate for each method, calculated from all 1300 picks. The second row provides the standard deviation of the success rate for each method across various scenarios. The first three columns of the table present an ablation comparison for our DYNAMO-GRASP (*DYN*) method, while *Dex* and *Cen* represent the DexNet and Centroid methods, respectively.

3.9.1 Large-scale, Diverse Scenario Assessment, and Ablation Test

To comprehensively assess the performance and robustness of various methods for the suction grasping challenge, we generated 260 diverse picking scenarios. We use the same simulation environment as we used to generate our training dataset. Each of the three methods was tested with five suction grasps per scenario in simulation, resulting in 1300 simulated suction grasps for each method's evaluation. The scenarios were generated by sampling from a distribution that incorporates even greater randomness in object orienta-

tion than the dataset used for model training. These scenarios incorporate a wide range of object configurations, leading to potentially complex object movements during picking.

Comparing the first, fourth, and fifth columns of Table.3.4, it is evident that our method exhibits a marked improvement over both DexNet and the Centroid method in terms of overall success rate and consistent performance across various scenarios. **Our method achieved the highest success rate of 88.05% and exhibited the least variance in success across different scenarios.** The first, second, and third columns of Table.3.4 presents an ablation test that illustrates the contributions of various components in our learning pipeline to the effective training of our model. *Dyn(Full)* is our final model, *Dyn w/ MSR* represents a model trained with standard MSR loss instead of the $\mathcal{L}_{Y_{max}}$ described in Sec.3.5.1, and *Dyn w/o PEN* further remove the use of penalization term p_{move} in the labeling process.

3.9.2 Real-world Evaluation

To assess real-world efficacy, we executed 375 real-world suction grasps to evaluate the various methods. In this experiment, we curated three sets of scenarios: the *Common set*, *Challenging set*, and *Adversarial set*, each embodying a distinct level or type of challenge for suction grasping. The statistics of all experimental trials and their comparison to the simulated trials are detailed in Table.3.1, 3.2, and 3.3.

The Common Set: In this experiment, we sampled ten scenarios from the 260 randomly generated ones as detailed in Sec.3.9.1. We then recreated these scenarios in the real world using objects with dimensions similar to those in the simulations. Each method was used to perform five grasps on each of these scenarios. This evaluation set captures the typical challenges of most picking tasks in this specific warehouse environment. As shown in Fig.3.7, **our model demonstrates an advantage with a total success rate of 94%, averaging 4.7 successful grasp out of five attempts and a standard deviation of 0.67**. In contrast, both DexNet and the Centroid method average 4.2 successful grasp out of five attempts. Their higher standard deviations, 0.92 and 1.03, point to less consistent performance.

The Challenging Set and Adversarial Set: We are particularly interested in the more challenging cases. Consequently, we devised two sets of scenarios in the real world to further test the capabilities of the three methods. The *Challenging Set* comprises five scenarios from the 260 scenarios described in Sec.3.9.1. These scenarios exhibit the lowest combined success rate for the three methods in simulation, representing the most challenging situations our simulation generated without human bias. In contrast, the *Adversarial Set* comprises five scenarios designed by a human operator, specifically tailored to challenge these grippers. The objects featured in this set are everyday items that were not included during the training phase. As depicted in Fig.3.7 and Table.3.2,3.3, **DYNAMO-GRASP markedly**

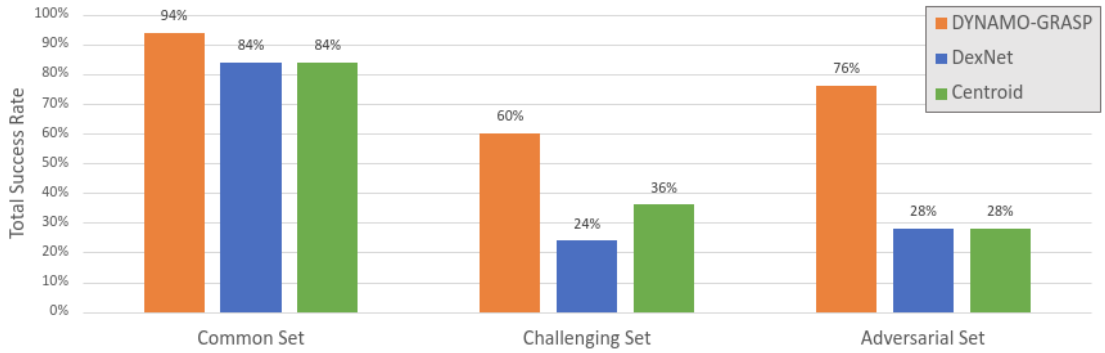

Figure 3.7: Comparison of the total success rates of different methods underscores their real-world performance on the three evaluation sets described Sec.3.9.2. The total success rate is computed by dividing the number of successful grasps by the total number of attempts within an evaluation set.

outperforms the two baseline methods in both total success rate and performance consistency in more challenging scenarios. On the challenging set, our method achieved a success rate of 60%, whereas, on the adversarial set, it reached 76%. In stark contrast, DexNet and the Centroid method's success rates are 24% and 36% for the challenging set, with both achieving 28% on the adversarial set. Furthermore, DYNAMO-GRASP consistently executed more than four successful grasps out of five attempts in over half of the scenarios in both sets. Meanwhile, the other two methods faltered, rarely managing even three successful grasps in any scenario within these evaluation sets.

Qualitative Analysis. Fig.3.6 depicts the grasp points chosen by various methods and indicates the success of each attempt during the adversarial evaluation. The figure offers insights into the areas chosen by each method for grasping and sheds light on which areas are more likely to lead to successful grasps. For example, in the first scenario, a tall bottle is partially propped

up by a box in the back. The test checks the grasp method's awareness of potential object toppling. DYNAMO-GRASP chose the bottle's lower part, ensuring the box supported the pick. Some grasp points chosen by the other two methods were higher up on the bottle, leading to toppling movements. Similarly, in scenarios two, four, and five, **DYNAMO-GRASP tends to select grasp points from regions that are overlooked by the other methods, resulting in more successful grasps in these scenarios.**

3.10 Conclusion and Limitation

This paper discusses the challenge of complex object movement during suction grasping, which no current state-of-the-art method adequately addresses. We introduced DYNAMO-GRASP, a dynamic-aware grasp point detection method that selects grasp points by factoring in the impact of object movement on the success of suction grasping. DYNAMO-GRASP delivers improved grasping performance with greater consistency in both simulated and real-world settings. Notably, in real-world experiments involving challenging scenarios, our method exhibits an improvement of up to 48% in success rate compared to alternative methods. **Limitations and future work:** Firstly, the dataset used in our simulation environment primarily includes objects with relatively simple geometric shapes. This aspect could limit the efficacy of our method when dealing with objects of uncommon or complex shapes. Similarly, our real-world experiments primarily involved simple ge-

ometric objects, such as boxes and bottles. In future research, there's potential to develop effective heuristics that combine information from both DYNAMO-GRASP and DexNet. While our method emphasizes modeling object movement, DexNet primarily targets suction quality based on object surface geometry. Integrating the strengths of both methods could lead to enhanced performance in specific applications.

3.11 DYNAMO GRASP RGB

We extended DYNAMO-GRASP to incorporate RGB image data, enhancing the system's capabilities. By integrating high-resolution RGB images of Google-scanned objects into the simulation environment, we refined grasping strategies. Domain randomization was applied to both the physical properties and visual appearances of objects, ensuring robust performance under varied real-world conditions.

Incorporating RGB data improved object recognition and grasp point detection accuracy. The enhanced visual input enabled DYNAMO-GRASP to predict object dynamics better, leading to more precise and reliable grasps. Evaluations in both simulated and real-world environments showed a 25% accuracy boost over the original DYNAMO-GRASP method. This extension highlights the potential of combining visual data with physics-based simulation to enhance robotic manipulation tasks.

Chapter 4

VLM-Based Zero-Shot Learning Grasp Method

4.1 Introduction

In the field of robotic manipulation, accurately identifying optimal grasp points on objects is crucial for the successful execution of tasks. Traditional methods often rely on extensive training datasets and specialized models to predict these grasp points. However, leveraging pre-trained models that have been designed for general tasks, such as Visual Language Models (VLMs) [55, 56, 57], can provide a more efficient and flexible approach. This chapter explores the application of zero-shot learning techniques using VLMs, specifically utilizing ChatGPT-4 [58], to predict the best grasp points for suction-based robotic grasping.

By employing prompt engineering, we aim to harness the capabilities of ChatGPT-4 [58] to identify the top candidate grasp points on an object from

a single-view depth image observation. This approach involves iteratively refining the grasp point predictions through a series of structured prompts and visual feedback. The goal is to enable robots to perform effective suction grasps even in challenging scenarios where the object may not have stable support or where precise positioning is required.

This chapter details the methodology and evaluation of using VLM-based zero-shot learning for grasp point detection, providing insights into the advantages and limitations of this novel approach.

4.2 Problem Statement

Our objective is to identify optimal grasp points on a target object within a container filled with multiple items using a single-view depth image observation. The identified grasp points should enable a robot to successfully establish a suction grasp, even when the object lacks stable support in the direction opposite to the robot's push. Consistent with previous suction grasp point detection studies, a grasp point is defined by a target point $[p, v]$. Here, $p \in \mathbb{R}^3$ represents the center of the contact ring between the suction cup and the object, while $v \in S^2$ denotes the gripper's approach direction. The grasp labeling function is defined as 1 if the grasp successfully forms a suction grasp on the target object and 0 if it does not. This section discusses the crucial criteria for forming a successful suction grasp.

To address the problem, we utilize Visual Language Models (VLMs) like

ChatGPT-4 [58] to predict the best grasp points using prompt engineering. The process involves:

1. **Initial Grasp Point Identification**: Using an RGB image annotated with potential grasp points and a corresponding depth image, the model is prompted to select the top 5 best grasp points. Each point is provided with a confidence score indicating its likelihood of success.

2. **Refinement through Visual Feedback**: The selected grasp points are further refined by presenting the model with images showing different suction cup projections on the target object. The model analyzes these projections to recommend the best grasp point, considering the entire surface area covered by the suction cup and ensuring the ring lies within the object boundaries.

3. **Final Grasp Point Selection**: The model performs another round of analysis by shifting the suction cup projections in four neighboring directions, further fine-tuning the grasp point based on the updated visual feedback.

The methodology leverages the generalization capabilities of VLMs to effectively predict grasp points without extensive retraining, aiming to improve the success rates of suction grasps in complex, real-world scenarios.

Figure 4.1: First prompt: To identify initial grasp points

4.3 Methodology

The methodology for the VLM-based zero-shot learning grasp method is structured around the use of prompt engineering with ChatGPT-4 [58]. This process involves multiple steps and iterations to identify and refine optimal grasp points accurately. The specific prompts used in this methodology are as follows:

4.3.1 Initial Grasp Point Identification

The first prompt is designed to identify the initial set of potential grasp points:

> *This is a vertical pod where the objects are placed vertically. The objects are resting on the bin's base, which is at the bottom. Objects are kept on the vertical pod on top of each other or on adjacent sides, and this view is the front view of the pod. And we are do-*

ing manipulation from the front side; also, the camera is facing the same side. I am trying to grasp the object using a vacuum-suction robotic gripper. The image is cropped with the bin and is annotated with numbers associated with each potential grasp point (representing the sample grasp point) on the RGB image. Also, a depth image is provided, which gives a reasonable estimation of the object's placement, which is the exact correspondence to the RGB image where the darker shades are in the near region, and light shades are in the far region. To successfully create a suction seal without deforming the suction cup a lot, also try to avoid edge areas of the object; give me the top 5 best grasp points on the object by specifying the annotated numbers provided on the RGB image, which will have a good suction seal. When providing the grasp point recommendations, provide the grasp point number, and also provide me with the confidence score for each point and how much it will have the probability of being a success. The target object size is around 9cm in height and 12cm in width, while the suction cup size is 4cm in diameter.

```
{
  "grasp_points": [
    {"point": [number], "reason": "NA"},
    {"point": [number], "reason": "NA"},
```

```
    {"point": [number], "reason": "NA"},
    {"point": [number], "reason": "NA"},
    {"point": [number], "reason": "NA"}
  ]
}
```

Replace '[number]' with the actual point numbers. Remember: No verbose output apart from the given json format

4.3.2 Refinement through Visual Feedback

The second prompt refines the initial grasp points by analyzing the visual feedback:

> *Here, I have provided 5 images that show different suction cup projections on other grasp points on the target object in the grid; each grid contains a cropped image of the target object. Tell me which point is the best grasp point out of these images, provided by analyzing the grasp area covered by the suction cup where the suction cup projection is marked with a white-colored ring, and ensure that the ring lies entirely on the image/object and not near the edges of the image/object. I want to pick an object with a weak-powered vacuum from a suction cup. The object's size is 9cm in height and 12cm in width, and the suction cup size is 4cm in diameter; now, please analyze again using this information. Please provide me with the row*

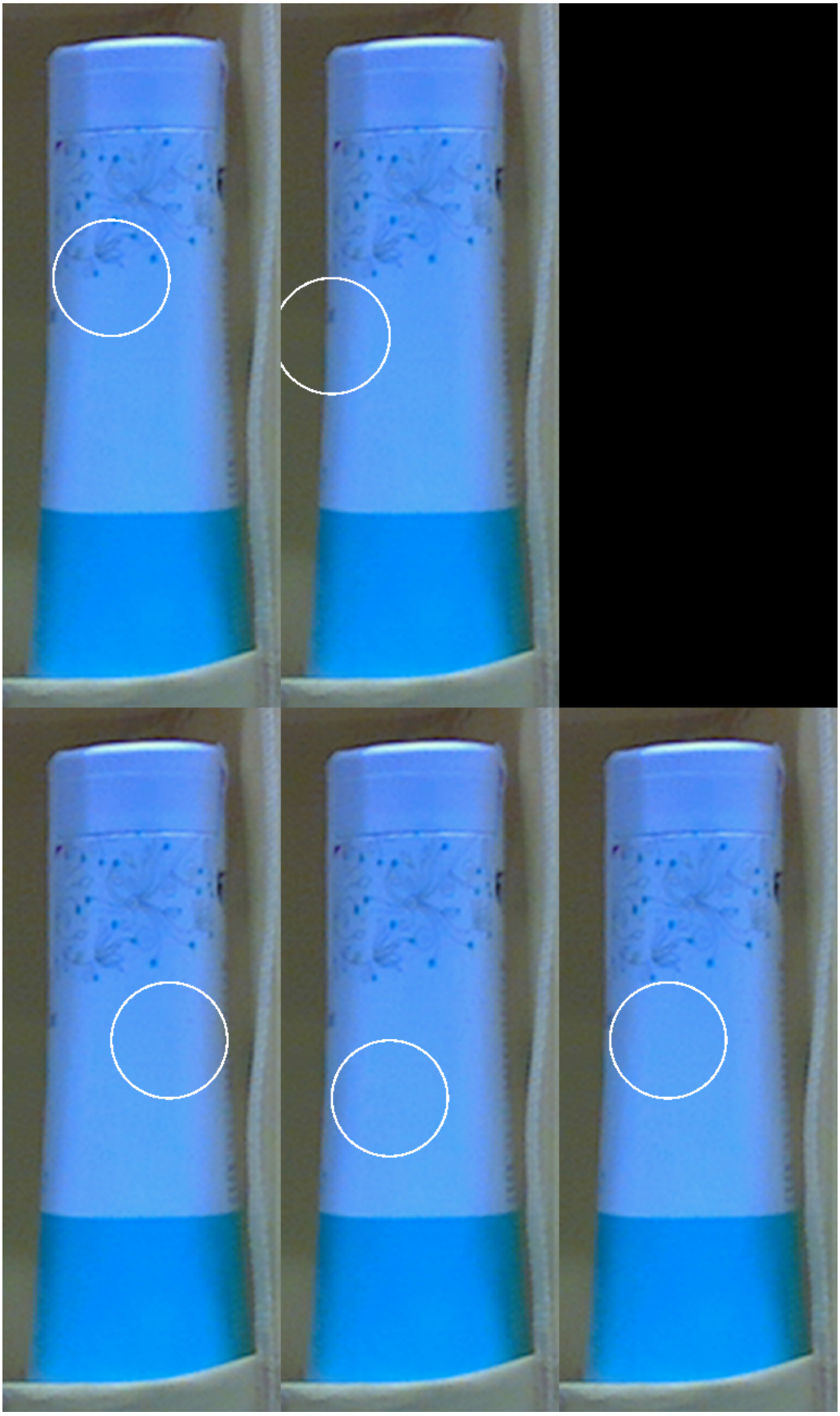

Figure 4.2: Second prompt: Annotate top 5 grasp points from the first prompt

and column number of the image where the image is concatenated in a 2x3 grid. There is no image in row 2 and column 3.

Provide the level of confidence for each grasp, such as:

- *Level 3 is the ideal scenario with a flat, rigid surface leading to consistent, successful grasps without issues.*
- *Level 2's grasp points near edges or corners present moderate challenges, resulting in a reduced but manageable success rate.*
- *Level 1 represents compromised grasp points at extreme edges or on flawed surfaces, where standard suction is largely unsuccessful and alternative methods are typically necessary.*

Give me a response as per the given format:

```
Row 1 Column 1: Level [X]
Row 1 Column 2: Level [X]
...
Best grasp point: Row [R] Column [C]
```

Replace 'X' with the level of confidence for each grasp point, and replace 'R' and 'C' with the row and column number of the best grasp point, respectively. Note: There is no image in row 2, column 3. Your analysis should be concise and strictly follow the given format. Remember: No verbose output apart from the given format.

Figure 4.3: Second prompt: Annotate grasp points, refining the grasp points by shifting the suction cup projections in four neighboring directions

4.3.3 Final Grasp Point Selection

The third prompt further refines the grasp points by shifting the suction cup projections in four neighboring directions:

> *Here, I have provided 5 images that show different suction cup projections on other grasp points on the target object in the grid; each grid contains a cropped image of the target object. Tell me which point is the best grasp point out of these images, provided by analyzing the grasp area covered by the suction cup where the suction cup projection is marked with a white-colored ring, and ensure that the ring lies entirely on the image/object and not near the edges of the*

image/object. I want to pick an object with a weak-powered vacuum from a suction cup. The object's size is 13cm in height and 7cm in width, and the suction cup size is 4cm in diameter; now, please analyze again using this information. Please provide me with the row and column number of the image where the image is concatenated in a 2x3 grid. There is no image in row 2 and column 3.

Provide the level of confidence for each grasp, such as:

- *Level 3 is the ideal scenario with a flat, rigid surface leading to consistent, successful grasps without issues.*
- *Level 2's grasp points near edges or corners present moderate challenges, resulting in a reduced but manageable success rate.*
- *Level 1 represents compromised grasp points at extreme edges or on flawed surfaces, where standard suction is largely unsuccessful and alternative methods are typically necessary.*

Give me a response as per the given format:

```
Row 1 Column 1: Level [X]
Row 1 Column 2: Level [X]
...
Best grasp point: Row [R] Column [C]
```

Replace 'X' with the level of confidence for each grasp point, and re-

place 'R' and 'C' with the row and column number of the best grasp point, respectively. Note: There is no image in row 2, column 3. Your analysis should be concise and strictly follow the given format. Remember: No verbose output apart from the given format.

By following these prompts, we systematically refine the grasp point selection process, leveraging the generalization capabilities of VLMs to identify the most suitable grasp points without extensive retraining. This methodology aims to improve the success rates of suction grasps in complex, real-world scenarios.

4.4 Results and Experiments:

For the experiments, a custom annotation tool was developed to select the area for the best grasp point on RGB images. This tool visualizes the suction cup on the image, allowing for precise annotation.

4.4.1 Data Collection

The data was collected in real-world settings, consisting of approximately 1000 object images and involved capturing RGB images of objects in a variety of settings:

- Deformable objects
- Objects with different sizes

Model Name	Accuracy
VLM zero shot	48.6%
Centroid	41.2%
DYNAMO-GRASP	30.2%
DYNAMO-GRASP RGB	59%
DexNet	41.2%
DexNet Fine-Tuned	47%

Table 4.1: Comparative evaluation of grasp success rates in for three methodologies: VLM zero shot model, DYNAMO-GRASP, DYNAMO-GRASP RGB, DexNet, DexNet Fine Tuned, and Centroid.

- Objects with different shapes

Each image was annotated using the custom tool to mark the best potential grasp points for a suction-based gripper.

4.5 Results

The results of the experiments are summarized in Table 4.1. The table compares the success rates for different object types and the overall performance of the VLM-based zero-shot learning grasp method. As you can see, the DYNAMO GRASP RGB model did a lot better than all the methods. VLM methods failed in cases where the bin was tightly packed, as it was difficult for the model to understand the information from the bin. Also, sometimes, it used to fail at random as the model is not deterministic; thus, it used to give failure results if run multiple times with the same inputs.

Chapter 5

GRASP-OPT: Optimized Grasp Detection using Synthetic RGB Images

5.1 Introduction

The advancement of robotic manipulation heavily relies on the development of models that can accurately identify and predict optimal grasp points on objects of various shapes and sizes. Recent research has emphasized the importance of integrating depth information and leveraging advanced neural network architectures to enhance grasp prediction accuracy [59, 21]. This chapter presents a novel approach that integrates a pre-trained Depth Anything Model with the DPT model and a custom Afford Grasp model. This combined model aims to enhance the performance of suction-based robotic grasping by providing detailed affordance maps that guide the robot in selecting the best grasp points and angles.

The Depth Anything Model is utilized for its capability to accurately esti-

mate depth information from RGB images [59]. Following this, the DPT model serves as the decoder, further refining the depth predictions [60]. The Afford Grasp model, acting as the head of the model, predicts three crucial affordances: the affordance map for the best grasp point, the roll angle, and the pitch angle. This comprehensive approach aims to significantly improve the accuracy and robustness of robotic grasping in diverse and complex environments. Predicting affordance map for suction grasping is very efficient along with orientation sampler [61].

Grasp optimization is critical for improving the success rates of robotic manipulation. Traditional approaches often involve extensive trial-and-error or rely on simplistic models that fail to generalize across different object types [62]. By leveraging advanced neural network architectures and integrating depth information, our proposed method addresses these limitations and provides a robust solution for grasp optimization.

The advancement of robotic manipulation heavily relies on the development of models that can accurately identify and predict optimal grasp points on objects of various shapes and sizes. This chapter presents a novel approach that integrates a pre-trained Dino V2 [63] model from the Depth Anything Model with the DPT (Dense Prediction Transformer) [60] model and a custom Afford Grasp model. This combined model aims to enhance the performance of suction-based robotic grasping by providing detailed affordance maps that guide the robot in selecting the best grasp points and angles.

The Depth, Anything Model [64] accurately estimates depth information from RGB images. Thus, we used pre-trained Dino V2 [63] from Depth Anything [64] for fine-tuning suction grasp affordances. Following this, the DPT model serves as the decoder, further refining the depth predictions. The Afford Grasp model, acting as the head of the model, predicts three crucial affordances: the affordance map for the best grasp point, the roll angle, and the pitch angle. This comprehensive approach aims to significantly improve the accuracy and robustness of robotic grasping in diverse and complex environments.

5.2 Problem Statement

The objective of this work is to develop a model that can predict optimal grasp points for suction-based robotic grasping. This model should identify the best grasp points and the appropriate grasp angles (roll and pitch) using RGB images as input. The predicted affordances should enable the robot to successfully grasp even in challenging scenarios with objects of varying shapes, sizes, and orientations.

Specifically, the model aims to predict: 1. An affordance map indicating the best grasp points on the object. 2. An affordance map for the roll angle required for the grasp. 3. An affordance map for the pitch angle required for the grasp.

GRASP OPT leverages the strengths of the Depth Anything Model for depth

estimation and the DPT model for depth refinement, combined with a custom Afford Grasp model to predict the necessary affordances for successful suction grasping.

5.3 Methodology

The proposed methodology integrates several components to achieve accurate and reliable grasp predictions. The overall architecture consists of a pre-trained Depth Anything Model, the DPT model as the decoder, and a custom Afford Grasp model acting as the head. The following steps outline the methodology:

5.3.1 Model Components

Depth Anything Model

The Depth Anything Model is a pre-trained neural network designed to estimate depth information from RGB images. It serves as the initial stage in our pipeline, providing a coarse depth map that represents the distance of objects in the scene from the camera.

DPT Model

Following the Depth Anything Model, the DPT [60] model is employed as a decoder. The DPT model refines the depth predictions, generating a more accurate and detailed depth map. This refined depth map is crucial for

accurately determining the grasp affordances.

Afford Grasp Model

The Afford Grasp model is a custom neural network designed to predict three specific affordances: 1. An affordance map for the best grasp point, indicating the most suitable locations on the object for a suction grasp. 2. An affordance map for the roll angle, which specifies the optimal rotation around the horizontal axis for the suction gripper. 3. An affordance map for the pitch angle, which specifies the optimal rotation around the vertical axis for the suction gripper.

The Afford Grasp model takes the refined depth map from the DPT model as input and outputs the three affordance maps.

5.3.2 Training and Data Collection

GRASP OPT is trained using a dataset consisting of RGB images, depth maps, and annotated affordance maps. The data includes various objects with different shapes, sizes, and materials to ensure robustness and generalization. The training process involves minimizing the loss function, which measures the discrepancy between the predicted affordance maps and the ground truth annotations.

5.3.3 Algorithm Workflow

The workflow of the proposed algorithm is as follows:

1. Input Processing: An RGB image of the scene is captured and passed through the pre-trained Dino V2 [63] from the Depth Anything Model to generate an initial depth map.

2. Depth Refinement: The initial depth map is refined using the DPT [60] model, resulting in a high-resolution depth map.

3. Affordance Prediction: The refined depth map is fed into the Afford Grasp model, which predicts the affordance maps for the best grasp point, roll angle, and pitch angle.

4. Grasp Execution: The predicted affordances are used to guide the robotic gripper to perform the suction grasp at the optimal location and orientation.

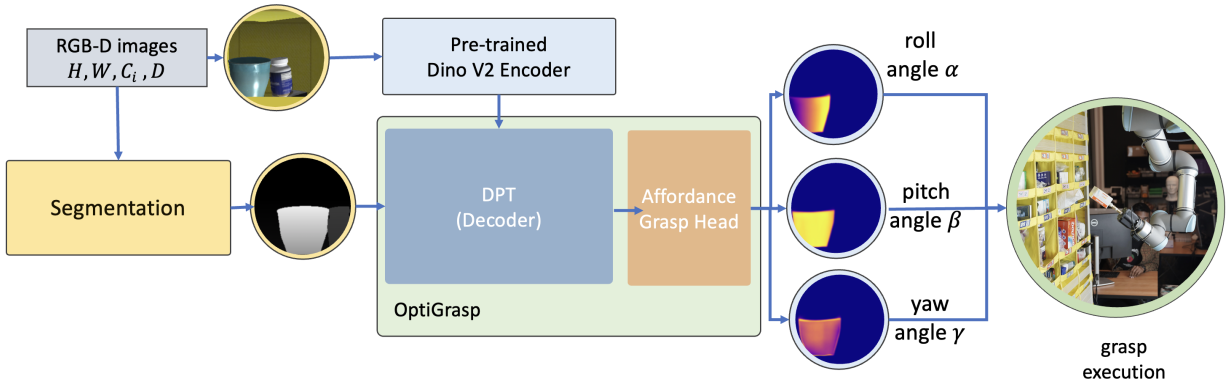

Figure 5.1: Architecture of the proposed model integrating Pre Trained DinoV2 from Depth Anything, DPT, and Afford Grasp models.

5.4 Results

The object sets were categorized into three levels of difficulty: easy, medium, and challenging.

1. **Easy Object Set**: This set primarily includes bottles and boxes. Although these objects are generally straightforward to grasp, their orientation can pose challenges. When placed in various orientations, identifying accurate grasp points can become difficult.

2. **Medium Object Set**: Similar to the easy object set, this set contains bottles and boxes but introduces additional complexity due to geometric irregularities. These irregularities can complicate grasping tasks. Furthermore, the medium object set includes objects with transparent sections. The training images did not encompass transparent or deformable objects, so the resulting affordance maps often lack consistency in these areas.

3. **Challenging Object Set**: This set comprises objects with minimal graspable areas or no valid grasp points, making them unsuitable for suction-based grasping due to hardware limitations. Additionally, the challenging set includes deformable objects, increasing the complexity of grasping tasks.

The performance of the proposed model is evaluated on 215 picks consisting of

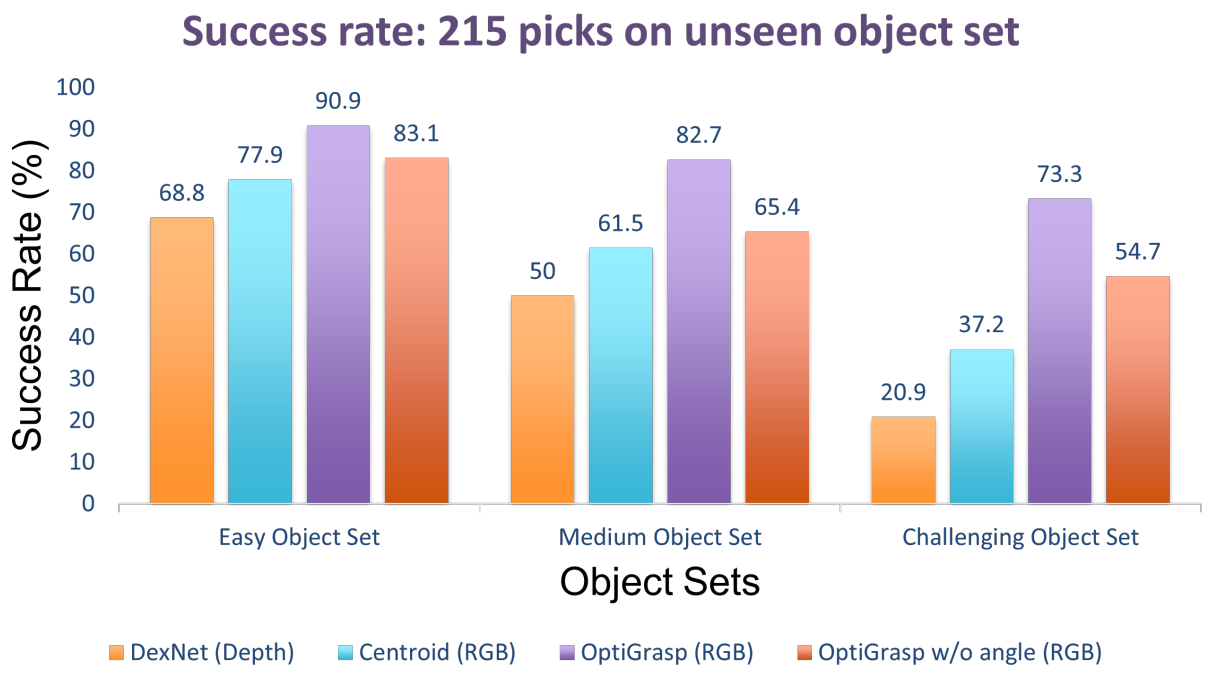

Figure 5.2: Success rate comparison between different baselines.

images of 170 unseen objects. The success rate of the suction grasps, and the accuracy of the predicted grasp points and angles are measured. Comparative studies are conducted to benchmark the proposed method against existing state-of-the-art techniques.

The results demonstrate the effectiveness of the proposed model in accurately predicting grasp points and angles, leading to a high success rate in robotic suction grasping tasks. The detailed affordance maps enable the robot to handle objects with complex geometries and varying sizes, highlighting the robustness and versatility of the approach.

On Synthetic data, the GRASP OPT method got 33% more accuracy than the baselines. Output on a few real-world objects is shown in 5.3 with a sim to the real transfer of GRASP OPT.

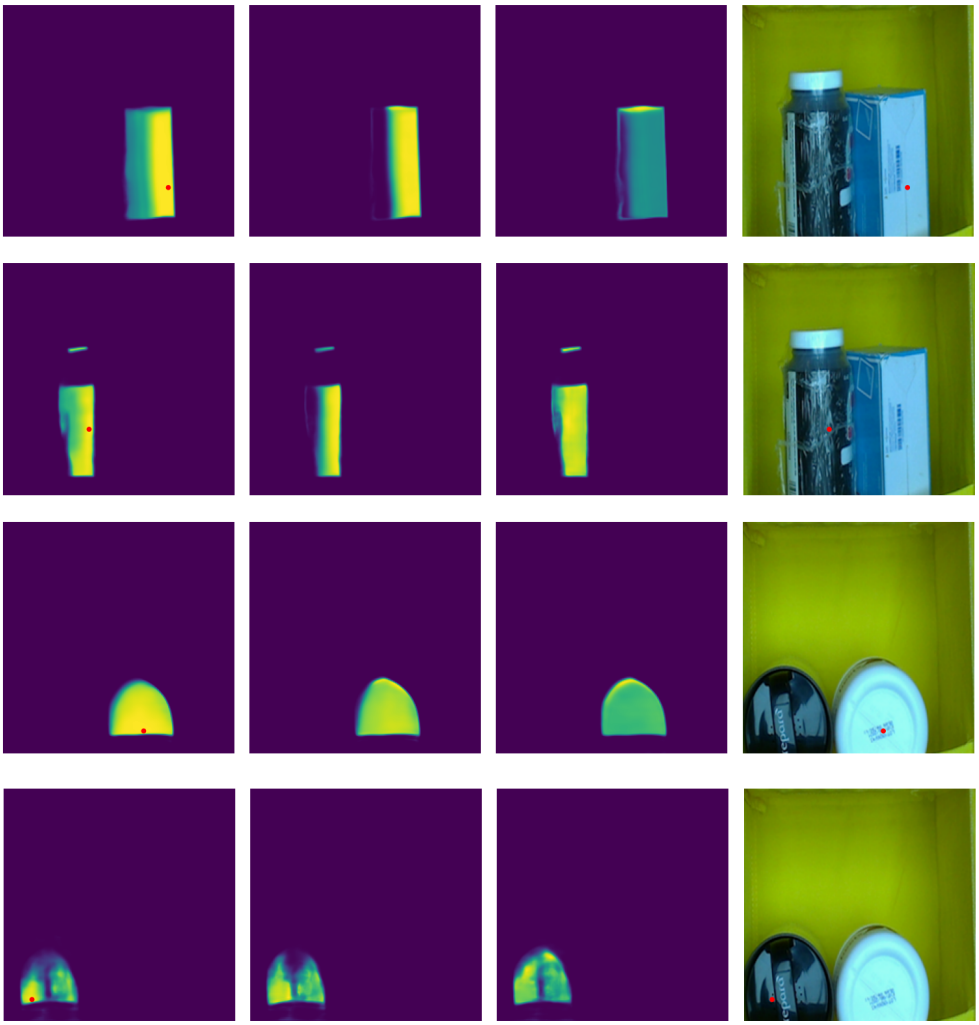

Figure 5.3: Visualization of grasp affordance maps and the RGB image for object manipulation. The first image depicts the grasp affordance map, highlighting the optimal grasp point with a red dot. The second and third images represent the affordance maps for roll and pitch angles, respectively, crucial for determining the grasp orientation. The fourth image is the RGB representation of the scene, with the red dot indicating the best grasp point on the object. These visualizations collectively aid in understanding and improving robotic grasping strategies.

Chapter 6

Future Directions

Looking ahead, one of the primary areas of focus will be the development of a Next Best View (NBV) algorithm to further enhance the efficiency and accuracy of robotic grasping tasks. The NBV algorithm aims to dynamically predict the optimal pose for the robot's end effector by considering the spatial configuration of objects within a bin. This approach is expected to improve the robot's ability to grasp objects by selecting the most advantageous viewpoint, thereby increasing the overall success rate of robotic manipulation in complex and cluttered environments. By integrating advanced sensing technologies and machine learning techniques, we aim to develop an adaptive system capable of real-time adjustments to grasping strategies based on continuous feedback from the environment.

In addition to the NBV algorithm, future work will also focus on improving computational efficiency to ensure quick and responsive operation of robotic systems in real-world scenarios. This includes optimizing existing algorithms and exploring the use of advanced hardware to accelerate processing times.

Furthermore, expanding the application domains of these robotic grasping technologies remains a key objective. Potential applications include enhancing warehouse automation, improving manufacturing processes, and advancing service robotics. By addressing these areas, future research can build on the foundations established in this thesis, contributing to the development of more sophisticated and versatile robotic systems capable of performing a wide range of tasks in diverse environments.

Bibliography

[1] Boling Yang et al. "DYNAMO-GRASP: DYNAMics-aware Optimization for GRASP Point Detection in Suction Grippers". In: *7th Annual Conference on Robot Learning*. 2023. URL: https://openreview.net/forum?id=_DYsYC9smK.

[2] Tongjia Zhang, Chengrui Zhang, and Tianliang Hu. "A robotic grasp detection method based on auto-annotated dataset in disordered manufacturing scenarios". In: *Robotics and Computer-Integrated Manufacturing* 76 (2022), p. 102329.

[3] Manman Yang et al. "A cooperative mobile robot and manipulator system (Co-MRMS) for transport and lay-up of fibre plies in modern composite material manufacture". In: *The International Journal of Advanced Manufacturing Technology* (2021), pp. 1–17.

[4] Albert S Olesen et al. "A collaborative robot cell for random bin-picking based on deep learning policies and a multi-gripper switching strategy". In: *Procedia Manufacturing* 51 (2020), pp. 3–10.

[5] Shun Hasegawa et al. "A three-fingered hand with a suction gripping system for warehouse automation". In: *Journal of Robotics and Mechatronics* 31.2 (2019), pp. 289–304.

[6] Max Schwarz et al. "Nimbro picking: Versatile part handling for warehouse automation". In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 3032–3039.

[7] Hannah S Stuart et al. "Suction helps in a pinch: Improving underwater manipulation with gentle suction flow". In: *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE. 2015, pp. 2279–2284.

[8] Hikaru Kumamoto et al. "Underwater suction gripper for object manipulation with an underwater robot". In: *2021 IEEE International Conference on Mechatronics (ICM)*. IEEE. 2021, pp. 1–7.

[9] Ping Yong Chua, T Ilschner, and Darwin G Caldwell. "Robotic manipulation of food products–a review". In: *Industrial Robot: An International Journal* 30.4 (2003), pp. 345–354.

[10] R Morales et al. "Soft robotic manipulation of onions and artichokes in the food industry". In: *Advances in Mechanical Engineering* 6 (2014), p. 345291.

[11] Kieran Gilday, James Lilley, and Fumiya Iida. "Suction cup based on particle jamming and its performance comparison in various fruit handling tasks". In: *2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE. 2020, pp. 607–612.

[12] C Blanes et al. "Technologies for robot grippers in pick and place operations for fresh fruits and vegetables". In: *Spanish Journal of Agricultural Research* 9.4 (2011), pp. 1130–1141.

[13] Seokhwan Jeong, Phillip Tran, and Jaydev P Desai. "Integration of self-sealing suction cups on the FLEXotendon glove-II robotic exoskeleton system". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 867–874.

[14] Tae Myung Huh et al. "A multi-chamber smart suction cup for adaptive gripping and haptic exploration". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021, pp. 1786–1793.

[15] Barbara Mazzolai et al. "Octopus-inspired soft arm with suction cups for enhanced grasping tasks in confined environments". In: *Advanced Intelligent Systems* 1.6 (2019), p. 1900041.

[16] Jared Nakahara, Boling Yang, and Joshua R Smith. "Contact-less manipulation of millimeter-scale objects via ultrasonic levitation". In: *2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*. IEEE. 2020, pp. 264–271.

[17] Jeffrey Mahler et al. "Dex-net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning". In: *2018 IEEE International Conference on robotics and automation (ICRA)*. IEEE. 2018, pp. 5620–5627.

[18] Hanwen Cao et al. "Suctionnet-1billion: A large-scale benchmark for suction grasping". In: *IEEE Robotics and Automation Letters* 6.4 (2021), pp. 8718–8725.

[19] Xavier Provot et al. "Deformation constraints in a mass-spring model to describe rigid cloth behaviour". In: *Graphics interface*. Canadian Information Processing Society. 1995, pp. 147–147.

[20] Ramesh Kolluru, Kimon P Valavanis, and Timothy M Hebert. "Modeling, analysis, and performance evaluation of a robotic gripper system for limp material handling". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 28.3 (1998), pp. 480–486.

[21] Jeffrey Mahler et al. “Learning ambidextrous robot grasping policies”. In: *Science Robotics* 4.26 (2019), eaau4984.

[22] Ping Jiang et al. “Learning suction graspability considering grasp quality and robot reachability for bin-picking”. In: *Frontiers in Neurorobotics* 16 (2022).

[23] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. “6-dof graspnet: Variational grasp generation for object manipulation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2901–2910.

[24] Andy Zeng et al. “Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching”. In: *The International Journal of Robotics Research* 41.7 (2022), pp. 690–705.

[25] Quanquan Shao et al. “Suction grasp region prediction using self-supervised learning for object picking in dense clutter”. In: *2019 IEEE 5th International Conference on Mechatronics System and Robots (ICMSR)*. IEEE. 2019, pp. 7–12.

[26] Matthew T Mason. “Mechanics and planning of manipulator pushing operations”. In: *The International Journal of Robotics Research* 5.3 (1986), pp. 53–71.

[27] Kevin M Lynch and Matthew T Mason. “Stable pushing: Mechanics, controllability, and planning”. In: *The international journal of robotics research* 15.6 (1996), pp. 533–556.

[28] Mehmet Dogar and Siddhartha Srinivasa. “A framework for push-grasping in clutter”. In: *Robotics: Science and systems VII* 1 (2011).

[29] Andy Zeng et al. “Learning Synergies between Pushing and Grasping with Self-supervised Deep Reinforcement Learning”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2018.

[30] Sergey Levine et al. “End-to-end training of deep visuomotor policies”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1334–1373.

[31] Andy Zeng et al. “Transporter networks: Rearranging the visual world for robotic manipulation”. In: *Conference on Robot Learning*. PMLR. 2021, pp. 726–747.

[32] Hongtao Wu et al. “Transporters with visual foresight for solving unseen rearrangement tasks”. In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 10756–10763.

[33] Chelsea Finn and Sergey Levine. "Deep visual foresight for planning robot motion". In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 2786–2793.

[34] Baichuan Huang et al. "Visual Foresight Trees for Object Retrieval From Clutter With Nonprehensile Rearrangement". In: *IEEE Robotics and Automation Letters* 7(1) (2022), pp. 231–238. DOI: 10.1109/LRA.2021.3123373.

[35] Rui Wang, Yinglong Miao, and Kostas E Bekris. "Efficient and high-quality prehensile rearrangement in cluttered and confined spaces". In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE. 2022, pp. 1968–1975.

[36] Clemens Eppner et al. "Lessons from the amazon picking challenge: Four aspects of building robotic systems." In: *Robotics: science and systems*. 2016, pp. 4831–4835.

[37] Carlos Hernandez et al. "Team delft's robot winner of the amazon picking challenge 2016". In: *RoboCup 2016: Robot World Cup XX 20*. Springer. 2017, pp. 613–624.

[38] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. *Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation*. 2022. arXiv: 2209.05451 [cs.RO].

[39] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. *CLIPort: What and Where Pathways for Robotic Manipulation*. 2021. arXiv: 2109.12098 [cs.RO].

[40] Samuel Li et al. *ShapeGrasp: Zero-Shot Task-Oriented Grasping with Large Language Models through Geometric Decomposition*. 2024. arXiv: 2403.18062 [cs.RO].

[41] Boling Yang et al. "Motivating physical activity via competitive human-robot interaction". In: *Conference on Robot Learning*. PMLR. 2022, pp. 839–849.

[42] Boling Yang et al. "Stackelberg Games for Learning Emergent Behaviors During Competitive Autocurricula". In: *arXiv preprint arXiv:2305.03735* (2023).

[43] Binghao Huang et al. "Dynamic Handover: Throw and Catch with Bimanual Hands". In: *arXiv preprint arXiv:2309.05655* (2023).

[44] Tao Chen, Jie Xu, and Pulkit Agrawal. "A System for General In-Hand Object Re-Orientation". In: *Conference on Robot Learning* (2021).

[45] Emanuel Todorov, Tom Erez, and Yuval Tassa. "Mujoco: A physics engine for model-based control". In: *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE. 2012, pp. 5026–5033.

[46] Erwin Coumans and Yunfei Bai. *PyBullet quickstart guide*. 2021.

[47] Viktor Makoviychuk et al. *Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning*. 2021.

[48] Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. "Better plain ViT baselines for ImageNet-1k". In: *arXiv preprint arXiv:2205.01580* (2022). URL: https://arxiv.org/abs/2205.01580.

[49] Yi Li et al. "STOW: Discrete-Frame Segmentation and Tracking of Unseen Objects for Warehouse Picking Robots". In: *7th Annual Conference on Robot Learning*. 2023.

[50] Miran Heo et al. *VITA: Video Instance Segmentation via Object Token Association*. 2022. arXiv: 2206.04403 [cs.CV].

[51] Bowen Cheng et al. "Masked-attention Mask Transformer for Universal Image Segmentation". In: 2022.

[52] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

[53] Markus Grotz et al. "Towards robustly picking unseen objects from densely packed shelves". In: *RSS Workshop on Perception and Manipulation Challenges for Warehouse Automation*. 2023.

[54] Kuan-Ting Yu et al. "A summary of team mit's approach to the amazon picking challenge 2015". In: *arXiv preprint arXiv:1604.03639* (2016).

[55] Shiyu Jin et al. *Reasoning Grasping via Multimodal Large Language Model*. 2024. arXiv: 2402.06798 [cs.RO].

[56] Reihaneh Mirjalili et al. *LAN-grasp: Using Large Language Models for Semantic Object Grasping*. 2023. arXiv: 2310.05239 [cs.RO].

[57] Chao Tang et al. *Task-Oriented Grasp Prediction with Visual-Language Inputs*. 2023. arXiv: 2302.14355 [cs.RO].

[58] OpenAI. *ChatGPT: May 2024 Version*. https://www.openai.com/. 2024.

[59] Pietro Mazzaglia, Taco Cohen, and Daniel Dijkman. "Information-driven Affordance Discovery for Efficient Robotic Manipulation". In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. 2024.

[60] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. *Vision Transformers for Dense Prediction*. 2021. arXiv: 2103.13413 [cs.CV].

[61] Ping Jiang et al. *Multiple-object Grasping Using a Multiple-suction-cup Vacuum Gripper in Cluttered Scenes*. 2023. arXiv: 2304.10693 [cs.RO].

[62] Jeannette Bohg et al. "Data-Driven Grasp Synthesis—A Survey". In: *IEEE Transactions on Robotics* 30.2 (2014), pp. 289–309. DOI: 10.1109/TRO.2013.2289018.

[63] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. 2023.

[64] Lihe Yang et al. *Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data*. 2024. arXiv: 2401.10891 [cs.CV].