WESTERN SYDNEY
UNIVERSITY

Machine Learning for Robotic Arm and Gripper Control

Rizwan Karim Rafat
22061272

A report submitted for
INFO7017 Postgraduate Project B
in partial fulfillment of the requirements for the degree of
Master of Data Science

Supervisor: Oliver Obst

School of Computer, Data and Mathematical Sciences
Western Sydney University

Spring, 2024

Abstract

Robotic grasping in unstructured and unseen environment is a complex challenge due to the difference in object shapes and orientations requiring advanced perception and adaptability for effective manipulation. This project explores and discusses the application of the Learning to Grasp (L2G) model, a machine learning-based approach, designed to predict stable grasp points in real time using 3D point cloud data. The L2G model uses feature extraction techniques to handle diverse objects, even under partial visibility which aims to improve robotic grasping performance in scenarios that lack predefined structures.

The primary objectives of this work were two: to verify the L2G model's performance in a simulated environment by evaluating grasp success rates, prediction accuracy and inference times and to attempt a real-world deployment of L2G on a physical robotic system, the Doosan A0509 robotic arm, to assess its practical viability. In simulation the L2G model shows promising results, achieving a high grasp success rate across a variety of object shapes and outperforms baseline models in both accuracy and efficiency. These simulation results suggests the model's potential for grasping tasks in changing and unstructured environments.

However during the physical setup and deployment on the Doosan A0509 the project encountered substantial challenges that limited full implementation. Extended setup time and configuration issues with network connectivity and compatibility between control systems hampered smooth operation. Communication instability between the Doosan control unit and the main computer caused delays while software limitations with the Dart Platform and DH Robotics software for the gripper further complicated synchronized control. These issues emphasized the need for more robust infrastructure to support real-world deployment of machine learning-based grasping models.

As future work, the project recommends adding 3D perception capabilities such as depth or stereo cameras to enhance the L2G model's object recognition and spatial awareness in real-world applications. Additionally future testing on alternative robotic systems such as the Franka Emika Panda would provide valuable insights into the model's generalization across platforms. These improvements could help advance the application of L2G allowing more adaptable and precise robotic grasping solutions for industrial and assistive environments. This work contributes to the field by demonstrating the potential of ML-based grasp prediction in unstructured environment while highlighting the technical considerations necessary for successful deployment on physical robotic systems.

Disclaimer: Generative AI was used to plan out the structure and format of this report in LaTeX as per the learning guide.

Contents

1 Introduction 1
1.1 Background 1
1.1.1 Context and Importance 1
1.1.2 Advancements in Grasp Prediction 1
1.1.3 The Learning to Grasp (L2G) Model 2
1.2 Problem Description 2
1.2.1 Challenges in Robotic Grasping in Unstructured Environments 2
1.2.2 Technical and Hardware Limitations 2
1.3 Project Objectives and Scope 3
1.3.1 Verification of the L2G Model in Simulation 3
1.3.2 Attempted Deployment on Doosan A0509 3
1.3.3 Future Work Considerations 3

2 Literature Review 4
2.1 ALOHA 2: An Enhanced Low-Cost Hardware for Bimanual Teleoperation 4
2.2 Collaborative Grasp Planning with Multiple Object Representations 4
2.3 Graspness discovery in clutters for fast and accurate grasp detection 5
2.4 ROS: an open-source Robot Operating System 6
2.5 Estimating a garment grasping point for robot 7
2.6 Learning Grasp Strategies with Partial Shape Information 7
2.7 Grasping of Unknown Objects using Deep Convolutional Neural Networks based on Depth Images 7

3 Methodology 8
3.1 Feature Extraction 8
3.1.1 PointNet++ Encoder 8
3.1.2 DeCo Encoder 8
3.2 Contact Point Sampling and Grasp Prediction 8
3.2.1 Differentiable Sampling for Contact Points 8
3.2.2 Grasp Regression 9
3.2.3 Grasp Ranking and Classification 9
3.3 Training Process 9
3.3.1 Data Preprocessing 9
3.3.2 Multi-Task Loss Function 9
3.3.3 Hyperparameter Tuning 9
3.3.4 Synthetic Datasets and Model Generalization 10

4 Experiments and Results 11
4.1 Simulation Results 11
4.1.1 Verification of L2G Performance 11
4.1.2 Evaluation of Model Generalizability 13
4.2 Real-World Deployment Limitations 13
4.2.1 Setup and Network Issues 13
4.2.2 Code Execution Challenges on Doosan 14

5 Doosan A0509 Setup 15
5.1 Physical Setup of the Doosan A0509 Arm 15
5.2 Hardware Connections and Power Supply 15
5.3 Software Installation and Configuration 17

6 Future Work 18
6.1 Complete Deployment on Physical Robots 18
6.1.1 Network and Hardware Compatibility Enhancements 18
6.1.2 Steps for Complete Deployment on Doosan A0509 18

6.2 Enhanced 3D Perception with Cameras 18
6.2.1 Benefits of Enhanced 3D Perception 18
6.2.2 Proposed Implementation of Camera Integration 19
6.3 Model Testing on Alternate Robotic Systems 19
6.3.1 Expected Benefits of Testing on Alternate Platforms 19
6.3.2 Challenges and Solutions for Multi-Platform Testing 19

7 Conclusion 20
7.1 Summary of Key Findings 20
7.2 Future Impact 20

List of Tables

1 Simulation-Based Success Rates for L2G and GPNet on ShapeNetSem-8 11
2 Inference Time Per Shape for L2G and GPNet 12
3 Generalization Results on YCB-76 for L2G and GPNet 13

List of Figures

1 Success-Coverage Curve for Grasp Prediction Models 12
2 Doosan A0509 Robotic Arm 15
3 Dooson Robotic Arm Control Unit 15
4 Unmanaged Network Switch 16
5 Gripper (DH-ROBOTICS PGC-140-50) 16
6 Aftermarket USB-to-serial converter 16
7 Variable Power Supply 17

1 Introduction

1.1 Background

1.1.1 Context and Importance

The field of robotic manipulation is important for advancing automation among different industries from manufacturing and healthcare to household robotics [1]. Robotic grasping involves the ability to pick up and secure objects which important for many applications. Despite it seems to be straightforward, grasping unknown or variable objects remains a complex task for robots e.g. in unstructured environments where objects differ in shape, texture, and orientation [2].

In a controlled industrial environment such as assembly lines and robotic manipulation has seen notable success. Robots are programmed to grab and move specific objects placed in known positions which allows a high degree of efficiency and reliability. However new applications in changing environments such as warehouses, healthcare settings and home environments introduce a layer of unpredictability. For example in a warehouse a robot must handle a variety of products in diverse orientations, often stacked or grouped closely [3]. Similarly, in healthcare applications, robots assisting in surgical settings need precise grasping capabilities to handle sensitive or irregularly shaped tools [4].

Household robotics, specifically in assistive roles for elderly individuals or those with limited mobility, represents another domain where advanced grasping capabilities are needed. Robots in these environments must handle everyday objects such as cups, books and utensils without predefined layouts while adapting to different lighting conditions, spatial configurations and object occlusions. This variation makes robust grasp prediction particularly challenging as successful models in structured environments often lack the adaptability required in these more changing environments [27].

1.1.2 Advancements in Grasp Prediction

Advancements in machine learning have played an important role in evolving robotic grasp prediction techniques from traditional rule-based methods to data-driven approaches. Early grasp prediction depended heavily on 2D vision systems where cameras captured flat two-dimensional representations of an object. These methods although effective in certain contexts but lacked depth information and failed to capture the full spatial geometry of objects and often resulted in unstable grasps [11].

The introduction of 3D perception using point cloud data has significantly improved grasp prediction accuracy and reliability. Unlike 2D images, 3D point clouds provide depth, shape and texture information which creates a comprehensive representation of the object. This allows machine learning models to learn precise grasp points, even in variable orientations and surface geometries [23]. Modern techniques in grasp prediction include the use of Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs) to process 3D point clouds that enables models to generalize to diverse objects while identifying stable grasps [5].

Synthetic datasets like Dex-Net [10] have further improved data-driven approaches in grasp prediction by providing a wide repository of labeled grasps on synthetic objects that allows models to train on a wide variety of potential scenarios. This extensive training data helps models predict successful grasps from different approaches even under uncertain or partially occluded conditions. Furthermore, deep learning architectures integrating object recognition with geometric feature extraction have shown notable improvements that helps robots operate more effectively in unstructured environments [6].

1.1.3 The Learning to Grasp (L2G) Model

The Learning to Grasp (L2G) model is a leading recent development in robotic grasp prediction. Unlike earlier models that often depend on assumptions about an object's position or gripper orientation that the L2G model is designed to predict six degrees-of-freedom (6-DOF) grasps using partial point cloud data. This capability allows it to make stable grasp predictions in cluttered environments where objects are frequently partially unknown [9].

L2G's technical approach involves complex feature extraction from 3D point clouds, combining local and global geometric features to make accurate grasp predictions. Using a combination of the PointNet++ [19] and DeCo [20] encoders, L2G processes fine-grained surface details and the overall object shape that allows the model to handle complex and unstructured scenes. L2G includes a differentiable sampling method that prioritizes contact points close to ground-truth positions that reduces error in grasp predictions. Additionally, it employs a multi-task objective function combining regression and classification to evaluate the predicted grasps and ensure both stability and feasibility [15].

By predicting 6-DOF grasps with minimum assumptions, L2G offers a flexible and adaptive solution suitable for real-world deployment in unstructured settings. Its use of partial point clouds allows it to maintain performance even in cluttered or partially blocked scenarios, marking it as a valuable model for applications that demand high adaptability and reliability.

1.2 Problem Description

1.2.1 Challenges in Robotic Grasping in Unstructured Environments

Robots in unstructured environments face significant challenges when grasping unknown objects. Unlike controlled setups, unstructured settings present objects in unpredictable orientations which often stacked with other items or partially masked. For example in a household environment a robot might need to handle kitchenware, books, or toys arranged in cluttered setting. Each object presents unique and different handling challenges which requires the model to consider the shape variations, size and surface texture.

Accurate grasp prediction in these settings is challenging without depth information as grasp models can misjudge the surface geometry of objects and result to an unstable grasps. Moreover the potential for visual noise such as shadows or changes in lighting complicates the grasping process, particularly in environments where objects are frequently moved or obstructed by other items.

Traditional grasp prediction models trained in structured settings struggle to generalize to these dynamic and unpredictable conditions. At the same time, advanced models like L2G that use both local and global geometric features are necessary to meet the demands of real-world applications.

1.2.2 Technical and Hardware Limitations

Beyond software challenges, technical and hardware constraints cause substantial barriers to successful deployment in real-world settings. Many traditional robotic systems are designed for specific tasks in controlled environments that rely on predefined object positions and fixed gripper configurations. However when robots are introduced to unstructured environments, they must adapt to unpredictable conditions without compromising performance [7].

Simulations, although are valuable for initial testing, they cannot fully replicate real-world conditions which leads to a gap between model performance in simulation and actual deployment. In this project the deployment of the L2G model on the Doosan A0509 robotic arm faced several technical issues including network setup challenges and communication reliability which hampered comprehensive testing. These limitations emphasized the complexities of deploying advanced grasp models on physical robots.

1.3 Project Objectives and Scope

1.3.1 Verification of the L2G Model in Simulation

The main goal of this project is to verify the L2G model's effectiveness within a simulated environment by focusing on grasp prediction accuracy and adaptability across different object configurations. By assessing L2G's ability to predict stable grasps in various scenarios this phase establishes a performance baseline for future real-world applications. The simulation enables detailed evaluation of the model's accuracy, robustness and speed confirming that L2G meets important benchmarks for potential deployment.

1.3.2 Attempted Deployment on Doosan A0509

A secondary goal is to deploy the L2G model on the Doosan A0509 robotic arm to examine its real-world performance. This phase aims to bridge the gap between simulation and practical application by configuring the hardware for live testing, setting up network connectivity and executing grasp predictions on physical objects. Despite facing technical challenges, particularly related to network reliability and integration the deployment attempt provides valuable insights into the practical requirements for implementing the L2G model on physical hardware [8].

1.3.3 Future Work Considerations

The limitations encountered in this project suggest several pathways for future work. Including depth or stereo cameras could enhance object perception which provides richer data and improving grasp prediction in unstructured environments. Additionally, testing L2G on alternative robots such as the Franka Emika Panda would evaluate the model's adaptability to diverse hardware, advancing its potential for wide application.

2 Literature Review

This section explores the current status of research in robotic grasping and manipulation. It covers in-depth examinations of the first four publications as well as summaries of the latter three, with a focus on advances, approaches, and issues in obtaining reliable grasps from 3D point clouds. The review aims to identify gaps and possibilities that our project intends to address, laying the groundwork for the replication and validation of the L2G technique. Notably, subsection 3.5 summarises a project completed with the Doosan A0509 robotic arm.

2.1 ALOHA 2: An Enhanced Low-Cost Hardware for Bimanual Teleoperation

The ALOHA 2 study [18], which is about developing low cost gear for two-person operation, is what motivated our project. The ALOHA 2 project wants to speed up research and data collection for fine-grained manipulation tasks by making the hardware designs public and giving full instructions and simulation models. The ALOHA 2 system builds on the ALOHA 1 system [18] by adding a number of important changes that are supposed to make the operation setting faster, easier to use, and more reliable. These changes are broken down into many parts, such as the support frame, cameras, simulation models, grippers, and gravity corrections.

For the grippers, the leader grippers have been redesigned from the original scissor design to a low-friction rail design. This makes the machinery simpler and decreases the pressure on the operators. This new design needs a lot less force to operate, which means that the user's hands will not get tired after long time operation. The follower grippers have also been improved with a low-friction design that decreases delay and boosts force output. Carbon fibre nylon is also used to improve the flexibility and longevity of the gripper fingers and supporting structures [18].

The passive gravity compensation mechanism takes the place of the original rubber band system. It makes the system last longer and keeps operators from getting tired by better managing the load during operation. This new design uses parts that are already available on the market and lets the user change the load balance for ease. The support frame is made simpler while still being rigid, which makes it easier for people and robots to work together. The new frame makes it easier to interact with the robots from different angles and lets bigger props be added [18].

New Intel RealSense D405 cameras have been added to the camera system. These cameras have a wider field of view, better resolution, and the ability to measure distance. Custom 3D-printed frames for these cameras reduce their size and the chance of collision, which makes teleoperation better, especially for jobs that need to be done with great precision and accuracy [18].

There is a detailed MuJoCo model of the ALOHA 2 workcell that can be used to collect realistic and scalable modelling data. This model makes it possible to precisely and accurately identify systems and learn and test policies better in simulated environment [18].

Overall, the ALOHA 2 project focuses on user comfort by using low-friction materials, parts that can be adjusted, and suggestions for regular breaks to reduce stress. The project also makes sure that it is strong by using simpler mechanical designs and long-lasting materials. Better grippers and gravity balancing systems make it possible to do difficult jobs with more accuracy and force. The ALOHA 2 project provides a strong, easy-to-use, and fast tool for studying two-handed manipulation [18].

2.2 Collaborative Grasp Planning with Multiple Object Representations

This paper discusses about how difficult it is to plan grasps in environments that are not structured, where robots may not be able to see or do things according to plan. Most of the time, traditional grasp planning methods use just one type of object representation, which might not be strong enough to deal with all the unknowns that are in real life. The authors suggest a model that blends different representations of an object to make grasp planning more reliable by using the best features of each representation [12].

The Task Space Regions (TSRs) [22] are an important linked work that deals with pose uncertainty in manipulator paths. In [23], the authors created probabilistic models for grasp generation, which use image and point cloud features to guess how well a grasp will work. In another paper [24] the authors worked on using human-demonstrated grasps to make systems more stable.

The suggested framework blends the outcomes of various grasp planning methods that used different types of input data. Planning based on object recognition and point clouds are the main parts of the system. Object recognition-based planning uses a library of known 3D models and picks pre-calculated grasps for recognised objects. It benefits from full object shape reasoning but depends a lot on how well the objects are recognised. Point cloud-based planning uses split point clouds to directly plan grasps without recognising objects. This gives us more options, but it is limited by only being able to see parts of objects at a time. The model puts these parts together using a confidence function that figures out the chance of a good grasp action based on the data that has been collected and how reliable each representation is [12].

Multiple object models are joined in the multi-planner framework to make it more stable. There is a confidence function for each model that shows how well it predicts grasp success based on the data that has been collected. The model chooses grasps that are likely to work based on multiple representations. This decreases the chance of mistakes caused by wrong scene analysis. The goal of this group effort is to get all the different grasp planning algorithms to agree on something. This will make sure that the end solutions can handle mistakes in sensing and sensor data processing [12].

The model was onboarded on the PR2 robot, which used a stereo camera to create point clouds. To test how well their method worked, the authors did studies with 25 common home items. The joint planner was tested in simulations against a naive planner that just takes the best result from the recognition method as true. When it comes to good grasps, the joint planner did better because it takes into account various representations. Tests done in real life on the PR2 robot showed that the joint planner worked better, especially when the objects were not in the database [12].

Finally, the suggested joint grasp planning model makes things more reliable in unstructured settings by mixing different representations of objects. This method deals uncertainties in awareness and performance well, which leads to better success rates in grasping tasks. According to the paper, using more than one source of information can make artificial grabbing much more reliable in real life [12].

2.3 Graspness discovery in clutters for fast and accurate grasp detection

This paper discusses and attempts to solve the problem of increasing grip detection accuracy and efficiency in noisy environments. Grip detection, a fundamental challenge in robotics, requires identifying possible grip sites on objects, particularly in unstructured environments. Traditional approaches frequently use uniform sampling for grasping, considering all points identically, which can be inefficient and incorrect due to the insufficient distribution of graspable locations in noisy environments [14].

The authors establish the idea of "graspness," a geometrically based quality metric for distinguishing graspable sections in noisy environments. The graspness score is calculated by analysing possible future grasp poses from a given location, which accurately predicts the chance of a successful grasp. This methodology enables a more specific selection of grip, considerably increasing the speed and accuracy of grasp detection approaches [14].

The suggested approach is a two-step procedure. The first step involves calculating graspness. The graspness score for each point is determined using a look-ahead search approach that considers the geometry of the surrounding environment. This procedure involves thoroughly evaluating potential future grasp positions from each place to determine its graspability. The second step is the creation of a graspness model. To make graspness calculations more realistic for real-time applications, the authors created a neural network-based graspness model. This model approximates the look-ahead search process by quickly forecasting point-wise graspness scores, often known as the graspable landscape [14].

Extensive tests were undertaken to confirm the graspness model's usefulness. The tests were carried out using the GraspNet-1Billion [25] benchmark, a large-scale dataset for generic object gripping. The findings showed that integrating the graspness model into current approaches improved grip detection accuracy and speed significantly. The graspness model was evaluated with a variety of object types, angles, and camera configurations. The model demonstrated consistent performance, showing its adaptability and wide application. The suggested method was compared to many advanced grasp detection algorithms. The findings showed a significant improvement in accuracy (more than 30 AP) and inference time, establishing the approach as a new standard for grab detection [14].

In conclusion, the paper makes a substantial contribution to grip posture identification by offering a focused method to finding graspable locations in noisy environments. The graspness idea and model offer a reliable, efficient, and accurate solution that outperforms existing approaches and establishes a new benchmark in the sector. This work advances robotic manipulation in unstructured situations, making it more dependable and effective.

2.4 ROS: an open-source Robot Operating System

The Robot Operating System (ROS) has become a robotics industry standard, offering a solid foundation for the development of complicated robotic systems. ROS was created to meet the increasing complexity and variety of robotic systems. Traditional frameworks frequently struggled with the problems provided by various hardware setups and the large codebase necessary for complex robotic activities. ROS came as a solution, focusing on peer-to-peer architecture, multilingual support, tool-based development, and a small, modular framework. The design ideas are based on the necessity for modularity and reusability. The peer-to-peer design enables seperate processes to interact effortlessly, eliminating obstacles common in central server architectures [26]. Multi-language support allows developers to work in their favourite programming languages, which improves cooperation and code reuse. It is a tool-based approach advocates a microkernel architecture, in which tiny, independent tools execute specialised functions, therefore decreasing system complexity [27].

ROS's basic characteristics include nodes, messages, topics, and services, which help to simplify communication between various sections of a robotic system. Nodes are autonomous processes that do computation and interact with each other via messages. One of ROS's most notable characteristics is its vast collection of tools for debugging, visualisation, and data logging. Tools like as rviz and rqt offer excellent visualisation capabilities, allowing developers to track sensor data, robot statuses, and other crucial information in real time [27].

ROS has had a substantial influence on robotics research and development by decreasing the entrance barrier for new researchers while also enabling quick prototyping and testing. Its open-source nature has resulted in a growing community that is constantly contributing to its development and growth. This collaborative environment has resulted in the development of various packages and libraries that cover a wide variety of robotic applications, including perception, navigation, manipulation, and human-robot interaction. ROS's modularity and versatility have made it a popular choice for educational purposes. ROS has been embraced by universities and research institutes throughout the world to teach robotics principles and perform research. Its vast documentation and active user community are useful tools for both learners and developers [27].

Despite its accomplishments, ROS has several obstacles. One of the main challenges is ROS' scalability for large-scale robotic systems. As the number of nodes and complexity of interactions grow, it becomes more difficult to manage communication and ensure real-time performance. Furthermore, while ROS has a wide range of development and debugging tools, novice users may find it difficult to integrate these tools into a coherent workflow.

ROS has transformed the field of robotics by offering a flexible, open-source framework for the construction of sophisticated robotic systems. Its emphasis on modularity, reusability, and community-driven development has made it an invaluable resource for academics, educators, and developers. As ROS evolves, it promises to drive additional advances in robotics, allowing for the development of more capable and complex robotic systems.

2.5 Estimating a garment grasping point for robot

In this paper [28], the authors described a method for determining the best grasping point for a single-arm robot to pick up a pile of garments. The primary goal was to centralise the gripped object for easy handling, particularly while loading washing machines. The authors suggested a sophisticated pipeline that uses depth and colour data from an RGB-D camera to calculate the optimal grabbing point. This method solves the necessity for a dependable and exact manner of managing clothes, which is a major difficulty in robotic manipulation.

The pipeline has several important phases. An AprilTag [29] is placed at the bottom of a laundry basket which determines if it is empty. If not, the system defines a Region of Interest (ROI) using fiducial markers [30] or point cloud segmentation with normal vectors. The depth data is then analysed to establish surface regularity. For uneven surfaces, the centroid of the crumpled item is designated as the grabbing point. For normal surfaces, colour information is used to project the point cloud into a 2D picture, which is then separated by colour using K-means clustering. The centroid of the biggest cluster is used as the gripping point. A specialised filter minimises the impact of shadows and wrinkles, boosting clustering accuracy [28].

Experimental findings demonstrate the usefulness of the suggested model, with the algorithm correctly recognising the grabbing point in 43 of 50 test cases. The most common faults observed were many objects with identical regions or colours that were difficult to differentiate. Testing on a physical robot revealed that gripping the centroid of the biggest cluster resulted in a more compact and centrally gripped object, boosting the chances of successful loading into a washing machine [28].

2.6 Learning Grasp Strategies with Partial Shape Information

This study addresses the difficulty of grasping unexpected objects in congested surroundings without a comprehensive 3D representation of the world. The proposed technique analyses the stability of various grasps based on noisy, incomplete shape information from depth sensors. Combining this with a kinematic description of a robot's arm and hand, the system calculates precise finger locations for optimal item grip. The authors tested their method on two robots using various manipulators, obtaining effective grasps on items of varying forms and appearances in both complex and simple environments. The algorithm was also used to unload objects from a dishwasher, confirming its usefulness and effectiveness in real-world circumstances [23].

The study is based on prior work [31], [32], [33], [34] and [35], utilising supervised learning to identify acceptable grip spots from 2D photos, but applies it to more complicated grasping challenges including multi-fingered hands and partial 3D data. The technique begins by extracting grip quality attributes from both 2D pictures and 3D point clouds, followed by supervised learning to predict successful grasp configurations. Extensive tests demonstrated that the system could successfully handle a wide range of unique items, even in highly complex surroundings [23].

2.7 Grasping of Unknown Objects using Deep Convolutional Neural Networks based on Depth Images

This study discusses a data-driven method to robotic grasping of unknown objects that use Deep Convolutional Neural Networks (DCNN) using depth photographs as the only input. This technique replicates a robot's visual perception by using pre-computed grasps from analytical planners along with rendered depth pictures for training. The DCNN generates grasping motions directly from depth information and is assessed using the KIT [37] and YCB [38] datasets as well as a large data grasping database, demonstrating high force-closure performance in both simulations and real-world tests on the ARMAR-III humanoid robot. The scalability and adaptability of the technique, which does not need hand-labeled data, suggest interesting applications for autonomous robotic grasping in complicated environment [36].

3 Methodology

In this section we discuss the approach used to verify the Learning to Grasp (L2G) model in simulation. Our primary goal was to validate the model's ability to predict accurate and stable grasps across diverse object types using simulated data without direct implementation on a physical robot. This section covers feature extraction, contact point sampling and training processes emphasizing techniques and tools optimized for the simulated environment.

3.1 Feature Extraction

3.1.1 PointNet++ Encoder

PointNet++ is a widely recognized framework for hierarchical feature extraction from point clouds which builds upon the original PointNet architecture [19]. It employs multi-scale grouping to cluster points based on proximity allowing the model to capture fine-grained surface details and inter-point relationships. For each grouped set of points the encoder performs a series of set abstraction operations which combines feature aggregation with max-pooling. This hierarchical approach produces robust scale-invariant representations of local geometry enabling the L2G model to identify key grasp features even in complex and irregular objects [39].

3.1.2 DeCo Encoder

DeCo (Denoise and Contrast) assists PointNet++ by focusing on global feature learning through graph convolutions and contrastive learning [20]. Operating in two branches, DeCo denoises local features and captures high-level structural cues enhancing model robustness against partial views and noisy data. Pre-trained on large synthetic datasets like ShapeNet, DeCo allows the L2G model to generalize among different objects and maintain accuracy even when certain features are masked. The use of DeCo contributes to L2G's ability to handle cluttered environments where only partial object data is available [9].

Together these encoders provide a comprehensive feature representation that balances local and global information which is important for reliable grasp prediction in a simulated setting. This robust feature extraction allows L2G to achieve high accuracy in grasp prediction regardless of object complexity or partial blockage.

3.2 Contact Point Sampling and Grasp Prediction

A central component of the L2G model is its ability to differentiate contact point sampling technique. This process allows L2G to select high-quality contact points from the point cloud and significantly enhance grasp prediction accuracy.

3.2.1 Differentiable Sampling for Contact Points

Unlike traditional grasping models that use either fixed or random contact points, L2G's sampling method is designed to be differentiable that allows it to adaptively select best and optimal contact points from the point cloud. Using a nearest-neighbor loss function, L2G aligns sampled points with ground-truth grasp points ensuring proximity to the actual contact areas of the object. This sampling strategy not only improves grasp reliability but also enables end-to-end training through backpropagation. This optimizes contact point selection for each unique object [40].

3.2.2 Grasp Regression

Following contact point sampling the model's grasp regressor computes a six degrees-of-freedom (6-DOF) grasp configuration which defines the gripper's approach relative to the object. This includes calculating yaw, roll and pitch angles with special attention to the pitch angle which makes sure that the gripper approaches perpendicularly to the surface and maximizes stability. The model predicts both the primary contact point and a secondary reference point which helps determine a stable grasp orientation among a wide variety of object geometries [41].

3.2.3 Grasp Ranking and Classification

L2G then ranks each predicted grasp based on its likelihood of success using a binary classifier which sorts grasps into feasible or infeasible categories based on surface alignment and stability factors. This classifier is trained to prioritize grasps that maximize contact stability using a binary cross-entropy loss function to optimize classification. By ranking grasps based on success probability L2G is able to create a refined list of high-quality grasp options which enhances its performance in complex simulated environments [17].

This grasp prediction pipeline allows the L2G model to dynamically adapt its grasp configurations to different object shapes and orientations. The model's flexible grasp prediction makes it a promising solution for future deployment in real-world environment.

3.3 Training Process

3.3.1 Data Preprocessing

Training data was generated using partial point clouds from ShapeNetSem-8 and YCB-76, which are widely used synthetic datasets in robotic manipulation research [38]. Each point cloud was normalized to fit within a unit sphere with random sampling applied to standardize the number of points. This ensures consistent input data which helps in achieving high training accuracy and efficient model generalization [10].

3.3.2 Multi-Task Loss Function

L2G's training process uses a multi-task loss function that simultaneously optimizes for contact point sampling, grasp regression and grasp classification. The sampling loss penalizes deviations of the sampled contact points from ground-truth positions ensuring that the model prioritizes points close to actual object contact areas. Grasp regression loss measures the Euclidean distance between the predicted and actual points with an additional angular penalty for orientation accuracy governed by a parameter λ that balances position and orientation precision [16]. Lastly the classification loss is implemented as a binary cross-entropy function by filtering infeasible grasps by training the model to recognize high-success grasps that further refines its prediction accuracy [42].

3.3.3 Hyperparameter Tuning

Key hyperparameters such as the number of sampled contact points (M) and neighborhood size (k) were optimized to balance performance and efficiency. An optimal configuration of M = 500 and k = 10 was selected to provide an efficient and accurate grasp selection. By tuning these parameters L2G achieved high grasp success rates without sacrificing computational efficiency which is very important for rapid inference in both simulated and real-time applications [13]. The model was trained using an NVIDIA Tesla V100 GPU with PyTorch which allowed for faster training on large point cloud data.

3.3.4 Synthetic Datasets and Model Generalization

To increase robustness the model was trained and validated on two synthetic datasets: ShapeNetSem-8 that features a variety of geometric shapes and YCB-76 which provides realistic object geometries found in common manipulation tasks [43]. This combination of datasets helps the model generalize to unseen objects and demonstrates effective cross-domain adaptability. Testing on YCB-76 also showed that L2G's predictions maintain accuracy even when applied to novel objects supporting the model's scalability for future applications.

4 Experiments and Results

This section discusses a detailed analysis of the L2G model's performance in simulation and the challenges encountered during attempted real-world deployment on the Doosan A0509 robotic arm. The simulation results include key performance metrics such as grasp success rates, inference times and generalization capabilities among different object shapes. The limitations faced in physical deployment highlight the complexities of integrating advanced machine learning models with existing robotic hardware.

4.1 Simulation Results

4.1.1 Verification of L2G Performance

To validate the L2G model's effectiveness in predicting robust grasps we conducted extensive simulations using the ShapeNetSem-8 dataset. This dataset contains a wide variety of objects with different shapes and complexities providing a comprehensive test setup for evaluating grasp success rates, prediction accuracy and inference times.

Grasp Success Rate:

- The grasp success rate is an important metric representing the percentage of grasp attempts that result in stable and secure grips. For each test object, L2G generates a set of possible grasp points and ranks them based on feasibility and stability.

- In simulations with ShapeNetSem-8, L2G achieved a success rate of 93.6% when considering the top 10% of predicted grasps. This high success rate demonstrates the model's ability to consistently identify the most stable and practical grasp points. In comparison, GPNet which is a well-regarded baseline model, recorded a success rate of 90% under similar conditions. Table 1 provides a detailed comparison of L2G's success rates among multiple thresholds (top 10%, 30%, 50%, and 100% of grasps).

- To test L2G's robustness, additional simulations evaluated success rates when considering broader sets of grasp predictions. At the top 30% threshold, L2G maintained a success rate of 90.1%, while GPNet showed a decline to 76.1% indicating a reduction in stability as more grasps were evaluated. This robustness across broader thresholds emphasizes L2G's consistent reliability, even in less optimal grasp setting.

Table 1: Simulation-Based Success Rates for L2G and GPNet on ShapeNetSem-8

Model	**Success Rate @10%**	**Success Rate @30%**	**Success Rate @50%**	**Success Rate @100%**
GPNet	90.0%	76.1%	72.3%	58.8%
GPNet*	92.2%	90.0%	82.3%	59.7%
L2G	**93.6%**	**90.1%**	**87.9%**	**82.0%**
L2G+DeCo	94.6%	93.5%	91.4%	82.9%

The Success-Coverage Curve (Figure 1) shows the relationship between the success rate and coverage among different grasp predictions in simulation. This curve is an important indicator of model performance as it shows how well each model maintains a high success rate as the coverage threshold increases. Coverage in this context represents the proportion of evaluated grasp candidates relative to the total set with higher coverage indicating that more grasps are included in the analysis.

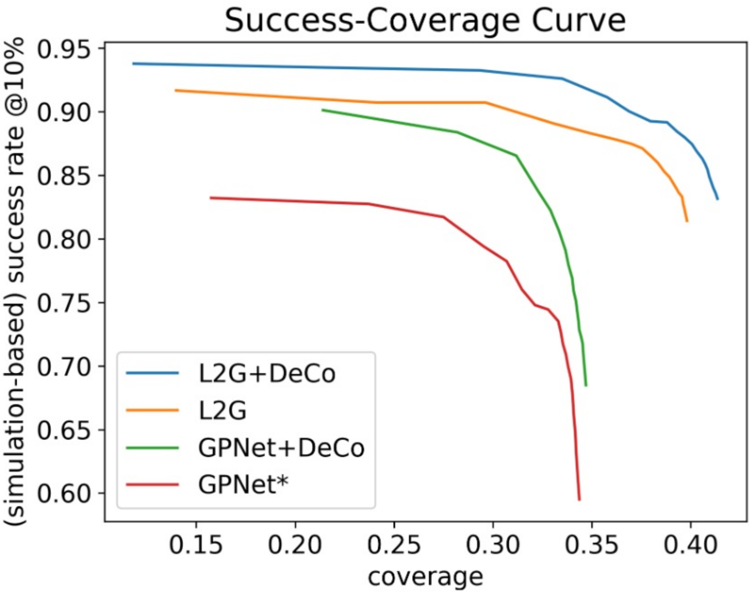

Figure 1: Success-Coverage Curve for Grasp Prediction Models

Prediction Accuracy:

- Prediction accuracy was measured by assessing L2G's ability to accurately identify feasible grasp points across a range of object shapes and orientations. Success rates at various thresholds provided a significant view of how well L2G ranks potential grasps and distinguishes stable grasps from less reliable ones.

- At the 10% threshold, L2G's accuracy remained high with a success rate of 93.6%, compared to GPNet's 90%. However, as the model assessed larger sets of grasps (e.g. the top 50% and 100% of predictions), L2G showed sustained performance achieving a success rate of 82.0% even at the top 100% while GPNet dropped to 58.8% at the same threshold. This comparison shows L2G's superior ability to prioritize high-quality grasp predictions among various object types and arrangements.

Inference Time:

- Efficiency in inference time is important for real-time applications. Table 2 presents a comparison of the inference times per shape for both L2G and GPNet showing that L2G required an average of 0.001 seconds at minimum and 0.339 seconds at maximum per grasp prediction. In contrast GPNet required 0.909 to 50.861 seconds per shape with its highest inference times occurring on complex geometries with high point counts.

- This efficiency advantage makes L2G more practical for real-world applications as it can process and predict grasps significantly faster than GPNet. The faster inference time is attributed to L2G's differentiable sampling technique and the computational optimizations within its architecture such as the use of DeCo for feature extraction which reduces processing overhead on complex point cloud data.

Table 2: Inference Time Per Shape for L2G and GPNet

Model	**Minimum Time (s)**	**Maximum Time (s)**
GPNet	0.909	50.861
L2G	**0.001**	**0.339**
GPNet+DeCo	0.935	46.897
L2G+DeCo	0.016	0.365

4.1.2 Evaluation of Model Generalizability

Evaluating L2G's generalizability is critical to understanding its performance on novel objects that were not included in the training data. To achieve this we tested L2G on the YCB-76 dataset which includes objects with diverse shapes and sizes to assess its capability to adapt to different object geometries.

Adaptability Across Object Shapes:

- L2G showed strong generalization abilities on the YCB-76 dataset maintaining success rates similar to those seen on the ShapeNetSem-8 dataset. Specifically L2G achieved a success rate of 45.0% at the top 10% threshold whereas GPNet showed a significantly lower success rate of 28.9%. This adaptability shows that L2G's feature extraction and sampling methods are robust across a variety of object types from simple geometries to complex asymmetrical shapes.
- L2G's performance on unseen objects confirms its capacity to generalize beyond the synthetic training dataset which is a key requirement for practical deployment in diverse real-world environments.

Robustness to Shape Variations:

- Further testing showed that L2G outperformed GPNet on the YCB-76 dataset by achieving higher coverage rates across all grasp thresholds. The DeCo encoder's contribution to global feature understanding helps L2G to adapt well to variations in object shape, orientation and complexity. L2G maintained an accuracy rate above 41.2% across all prediction thresholds which indicates robust performance even when lower-quality grasps were evaluated as shown in Table 3.

Table 3: Generalization Results on YCB-76 for L2G and GPNet

Model	**Success Rate @10%**	**Success Rate @30%**	**Success Rate @50%**	**Success Rate @100%**
GPNet	28.9%	29.2%	27.2%	20.8%
L2G	**45.0%**	**44.6%**	**43.8%**	**41.2%**

4.2 Real-World Deployment Limitations

Although L2G showed impressive results in simulation, our attempts to deploy the model on the Doosan A0509 robotic arm faced significant obstacles. These challenges included the complexities involved in transferring high-performing models from simulation environments to physical hardware.

4.2.1 Setup and Network Issues

Network Configuration Challenges: The setup process for the Doosan A0509 involved establishing a stable connection between the robot and the primary control computer. IP configuration was an initial challenge as IP conflicts prevented consistent communication between the devices. Multiple reconfigurations and consultations with network specialists were required to maintain connection stability.

Firewall and Security Protocols: Security protocols on the network firewall also interfered with Doosan's ability to receive control commands resulting in disconnections. Despite adjustments to firewall settings, stability remained an issue and network disconnections continued to occur unexpectedly, limiting effective real-time control of the robot.

4.2.2 Code Execution Challenges on Doosan

Despite resolving some network-related issues executing the L2G model on the Doosan A0509 presented further more difficulties due to code compatibility constraints and hardware limitations.

Insufficient Onboard Resources: The Doosan's onboard resources were inadequate to process the high-dimensional computations required by L2G. Attempts to offload computations to an external GPU server introduced latency, interrupting real-time processing and ultimately making the setup infeasible.

Dependency Conflicts and Software Limitations: L2G relies on specific libraries optimized for NVIDIA GPUs but these were incompatible with the Doosan's processing environment. Even with alternative configurations maintaining the model's accuracy proved challenging preventing successful execution on the robot.

These challenges show the importance of carefully aligning computational resources and software environments to achieve a smooth deployment of advanced models in real-world applications.

5 Doosan A0509 Setup

The setup of the Doosan A0509 robotic arm involved configuring the physical, network and software components to enable precise control of the robotic arm and gripper. The installation process included securing the arm, setting up network connectivity, configuring Windows 11, and installing the Dart Platform and DH Robotics software.

5.1 Physical Setup of the Doosan A0509 Arm

Mounting the Robotic Arm: The Doosan A0509 robotic arm (See Figure 2) was securely bolted onto a sturdy table to provide stability during operations. The table setup prevented any potential vibrations or displacement during tasks which was essential to maintain the accuracy and repeatability of the robotic arm's movements.

Control Unit Placement: The robotic arm is connected to a dedicated control unit (See Figure 3) which manages all operational commands. The control unit was positioned close to the arm, minimizing the cable length and reducing possible latency. This proximity is important for maintaining real-time communication between the arm and the control unit during complex movements.

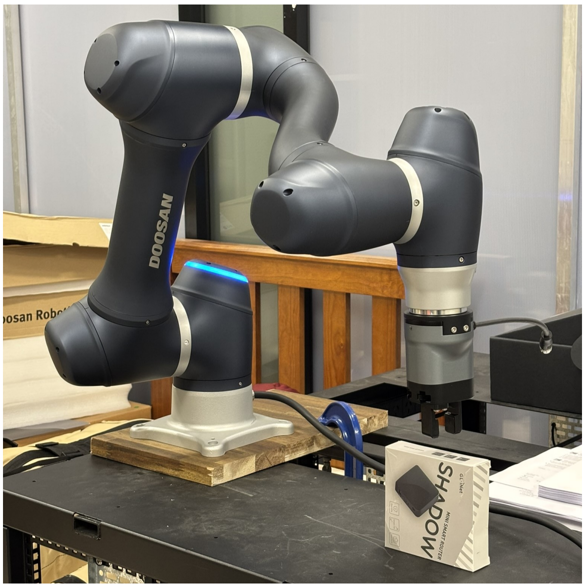

Figure 2: Doosan A0509 Robotic Arm

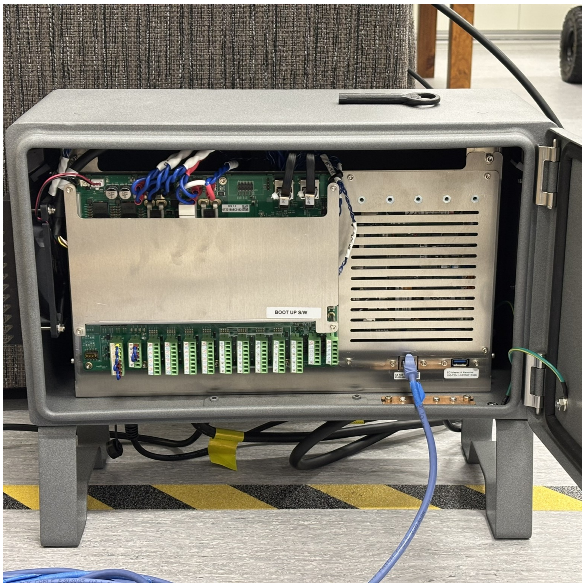

Figure 3: Dooson Robotic Arm Control Unit

5.2 Hardware Connections and Power Supply

Network Connectivity: To facilitate communication between the Doosan control unit and the main computer an unmanaged network switch (See Figure 4)[1] was used. The control unit was connected to the network switch, which in turn connected to the project computer. This setup enabled a stable and high-speed network connection between the components allowing for efficient data transfer during operation.

[1] Available at: https://www.netgear.com/au/business/wired/switches/unmanaged/gs105/

Figure 4: Unmanaged Network Switch

Direct Gripper Connection: The gripper (See Figure 5), an essential component for handling grasping tasks, was connected directly to the main computer via a USB-to-serial converter (See Figure 6). This connection allowed the DH Robotics software to interface with the gripper independently from the Doosan control unit, providing precise control over the gripper's actions.

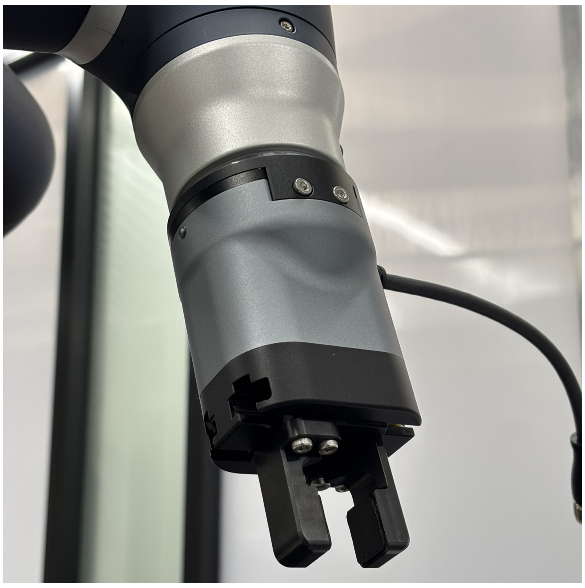

Figure 5: Gripper (DH-ROBOTICS PGC-140-50)

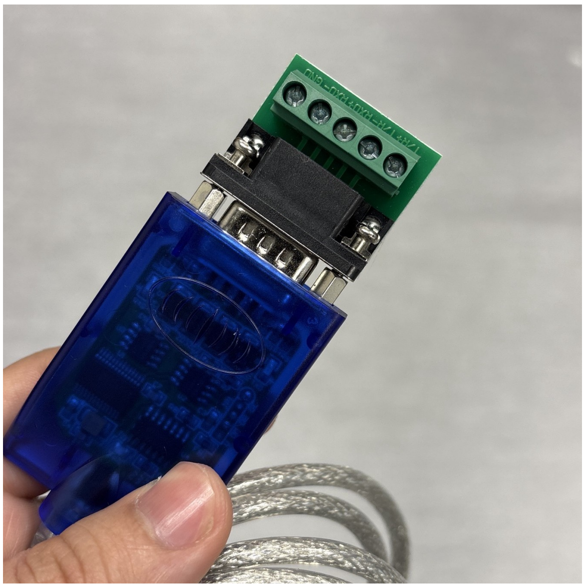

Figure 6: Aftermarket USB-to-serial converter

External Power Supply for the Gripper: The gripper required an external 24V power supply, which was provided by a dedicated power source (See Figure 7). This setup ensured that the gripper received consistent power independently of the robotic arm's control unit, reducing the risk of power fluctuations affecting the stability of the overall system.

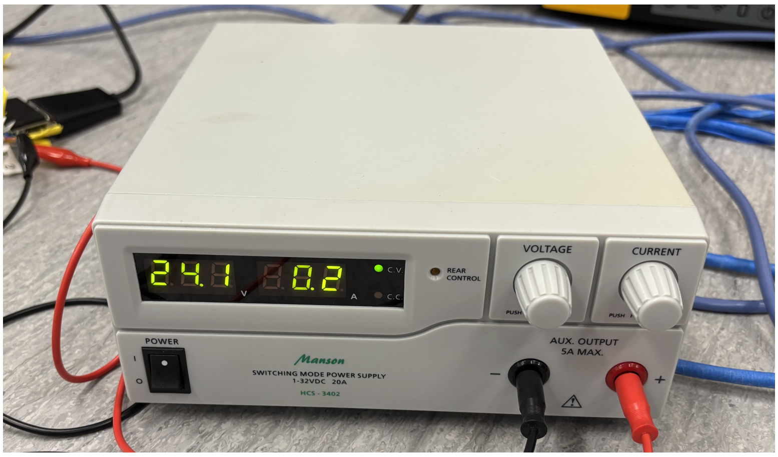

Figure 7: Variable Power Supply

5.3 Software Installation and Configuration

Operating System: The project computer was set up with Windows 11 as the operating system. Windows 11 was chosen for its compatibility with the Dart Platform[2] and DH Robotics software[3], both essential for controlling the Doosan A0509 robotic arm and gripper.

Dart Platform for Arm Control: The Dart Platform software was installed on Windows 11 to manage the robotic arm's movements. Dart provides a graphical interface and command-line tools for configuring and controlling the Doosan A0509. This platform allows the programming of movement sequences and for remote monitoring and control providing robust support for various robotic tasks.

DH Robotics Software for Gripper Control: The gripper was managed using DH Robotics software which provided precise control over the gripper's opening, closing and force settings. This software connected directly to the gripper via the USB-to-serial converter allowing operators to control gripper actions independently of the arm movements. The DH Robotics software was compatible with Windows 11 enabling smooth integration with the overall system setup.

Network Configuration: The control unit and the project computer were configured with static IP addresses within the same subnet for continuous communication. Initial IP conflicts were resolved by setting dedicated IP addresses to avoid overlap with other devices on the network allowing for a stable connection between the robotic components.

[2] Available for download only after successful sign-up and login at: https://www.robotlab.doosanrobotics.com

[3] Available for download at: https://en.dh-robotics.com/service/software

6 Future Work

6.1 Complete Deployment on Physical Robots

Despite extensive testing in simulation, the deployment of the Learning to Grasp (L2G) model on the Doosan A0509 robotic arm faced network stability and compatibility challenges that hampered full implementation. Future efforts should focus on addressing these challenges to ensure successful operation in physical environments.

6.1.1 Network and Hardware Compatibility Enhancements

A primary recommendation is to improve network reliability for the Doosan A0509 setup. Setting up a dedicated, isolated network specifically for the robotic system would minimize data transmission delays and eliminate IP conflicts caused by other devices on shared networks. A high-speed Ethernet connection with Quality of Service (QoS) protocols will further enhance performance by prioritizing data packets essential for real-time control.

In addition, rigorous testing of hardware compatibility across different configurations would be beneficial. Ensuring that all components including the control unit, network interfaces and gripper power supply are compatible with the robotic arm's control system is important for seamless integration. Using robust power supplies and shielded cables can prevent signal interference which is often a source of connectivity issues in industrial robotic setups.

6.1.2 Steps for Complete Deployment on Doosan A0509

To complete the deployment detailed documentation of the setup process, network configuration and troubleshooting protocols is recommended. By establishing a step-by-step guide future operators could replicate the setup more accurately thereby reducing downtime and minimizing errors in configuration. This documentation would serve as a foundational reference particularly when multiple operators are involved in system setup and testing.

Furthermore, exploring alternative integration software compatible with Windows 11 may provide a more reliable platform for communication between the robotic arm and control unit. If feasible, moving from a split control system (Dart Platform for the arm and DH Robotics software for the gripper) to a unified platform would allow for more synchronized control by simplifying the deployment process.

6.2 Enhanced 3D Perception with Cameras

Integrating depth or stereo cameras can greatly enhance the L2G model's performance by providing additional environmental data by enabling the robot to perceive its surroundings with greater accuracy and reliability.

6.2.1 Benefits of Enhanced 3D Perception

Enhanced 3D perception using depth cameras (e.g., Intel RealSense D435 or Azure Kinect) or stereo cameras allows the robot to capture detailed depth information that helps the point cloud data used in grasp prediction. By improving spatial awareness these sensors would allow the L2G model to better identify object boundaries, contours and blockages which are important for successful grasping in cluttered environments.

For instance, depth sensors could provide real-time depth mapping which would improve the arm's ability to distinguish between overlapping objects or objects with similar shapes. Stereo cameras with their ability to capture depth from two perspectives would further enhance the model's grasping capabilities by providing richer point clouds and improving the detection of object orientation and position.

6.2.2 Proposed Implementation of Camera Integration

Incorporating camera data into the existing L2G model would require additional software and hardware modifications. Specifically, integrating data from stereo or depth cameras would help modify the model's feature extraction pipeline to include depth features alongside the 3D point clouds. This additional data could be processed using convolutional neural networks (CNNs) or other vision-based methods designed to handle 3D input. These enhancements would improve the robot's grasping ability in variable lighting conditions and on textured or complex surfaces.

Testing different configurations such as positioning multiple cameras around the robotic workspace would allow more comprehensive coverage of the operational environment. By evaluating the grasp success rates across these configurations the project could determine the optimal placement and calibration settings to maximize the efficacy of 3D perception enhancements.

6.3 Model Testing on Alternate Robotic Systems

Testing the L2G model on different robotic platforms such as the Franka Emika Panda would provide valuable insights into the model's adaptability and generalizability across various hardware systems. This step is important for assessing the model's real-world applicability beyond the Doosan A0509.

6.3.1 Expected Benefits of Testing on Alternate Platforms

The Franka Emika Panda robotic arm, known for its high precision and adaptive control capabilities presents an ideal test platform to evaluate the generalization capabilities of the L2G model. Testing on an alternative system would demonstrate whether the model's grasp predictions and adaptability hold across different types of robotic hardware. If successful, this validation would establish the model's versatility for use in diverse industries including manufacturing, logistics and healthcare.

Deploying the L2G model on alternative robots would also highlight any platform-specific limitations or dependencies within the model's code. For example, differences in end-effector design, joint configuration and motor precision between the Doosan A0509 and the Panda arm might reveal new insights into the L2G model's performance under varying physical conditions.

6.3.2 Challenges and Solutions for Multi-Platform Testing

One of the primary challenges in multi-platform testing is ensuring that the model's control commands are compatible with the target robot's software and hardware interface. The Franka Emika Panda utilizes the Franka Control Interface (FCI) and Franka ROS which may require code modifications or middleware integration to accommodate L2G's grasp planning algorithms. Developing an abstraction layer that translates L2G's output into commands compatible with different robot interfaces would be essential for seamless cross-platform functionality.

Additionally each robotic platform may have different computational requirements. For example, the Panda arm's control system is optimized for its proprietary ROS-based architecture which could help adjust to the L2G model's processing pipeline to achieve real-time performance. Future work should explore optimizing the model for efficient inference on various platforms perhaps by including lightweight neural networks or alternative computational frameworks such as TensorFlow Lite.

7 Conclusion

7.1 Summary of Key Findings

This project aimed to verify the effectiveness of the Learning to Grasp (L2G) model in robotic grasp prediction focusing on its simulation-based performance and exploring deployment challenges on the Doosan A0509 robotic arm. The primary objectives were to validate the model's grasp success rate, accuracy and inference time in simulation and to implement the model on physical hardware for real-world testing. In simulation the L2G model demonstrated strong performance achieving a high grasp success rate across diverse object geometries. The model consistently outperformed baseline algorithms such as GPNet by maintaining higher prediction accuracy even on complex shapes and demonstrating efficiency in inference time across multiple grasp thresholds. This robust simulation-based verification confirmed the model's potential as an effective tool for robotic manipulation tasks in unstructured environments.

Despite this success, deploying L2G on the Doosan A0509 encountered several technical challenges that hampered full real-world testing. Network connectivity between the robotic control unit and the main computer faced stability issues due to IP conflicts, firewall restrictions and limited data transmission speeds leading to delays and occasional communication failures. The software architecture presented additional obstacles as the Dart Platform used to control the Doosan arm and the DH Robotics software for the gripper lacked full compatibility for synchronized operation. At the same time, these issues limited the system's real-time responsiveness and prevented seamless integration of the robotic arm and gripper for synchronized grasping actions. These findings emphasize the importance of robust network configurations, compatible control systems and synchronized software platforms for real-world deployment.

The contributions of this project extend across several dimensions in robotic grasping research providing insights into the effectiveness of machine learning models for grasp prediction as well as the limitations and challenges of deploying such models in physical settings. The simulation-based verification of L2G offers significant evidence supporting the model's adaptability and reliability in handling diverse object shapes and unstructured scenarios which are important requirements for advanced robotic manipulation. The simulation results demonstrate that machine learning models when adequately trained on synthetic datasets and equipped with feature extraction methods like PointNet++ and DeCo can achieve high accuracy in grasp prediction. This success highlights the L2G model's suitability for dynamic environments where traditional rule-based models may fall short due to their lack of adaptability to variable object geometries.

In addition to validating L2G's performance this project clarifies on the technical challenges inherent in translating ML-based grasping models from simulation to real-world deployment. Issues encountered during the Doosan A0509 setup reveal that while ML models can be highly effective in controlled simulation environments, their deployment on physical robotic systems is far from straightforward. The need for stable network infrastructure, compatible software and cohesive control interfaces remains a critical factor for real-world implementation. This work has thus contributed not only to verifying the effectiveness of L2G but also to identifying practical limitations that researchers should consider when planning similar implementations. These insights are valuable for the robotics and machine learning communities, as they highlight potential areas for infrastructure improvements, particularly in industrial settings where network reliability and system compatibility are crucial.

7.2 Future Impact

Looking forward the successful deployment of the L2G model with enhanced 3D perception technologies, such as depth or stereo cameras and testing on alternative robotic platforms like the Franka Emika Panda could significantly advance the field of robotic grasping. Integrating depth sensors would provide the L2G model with richer environmental data allowing it to adapt more effectively to complex, cluttered and partially obscured scenarios. By enhancing the model's spatial awareness and object identification accuracy, such additions would make it possible to deploy the model in applications where robots must interact with diverse objects under challenging conditions such as warehouse automation, healthcare assistance and home robotics.

Testing L2G on alternative robots would offer further insights into its ability to generalize and robustness across different hardware configurations opening up possibilities for multi-platform applicability. This capability is essential for the widespread adoption of ML-based grasp prediction models as industries rely on a range of robotic systems to meet diverse operational requirements. The ability of L2G to generalize across platforms would validate its use as a flexible and adaptable solution for advanced robotic manipulation tasks further expanding its relevance in fields that demand high adaptability, such as automated assembly, logistics and remote assistance.

In conclusion, this project has provided considerable evidence for the viability of ML-based models like L2G in grasp prediction while also highlighting the important role of supporting infrastructure in real-world deployment. By addressing the identified challenges and continuing to enhance the model with complementary perception technologies and cross-platform testing, future research can push the boundaries of robotic grasping bringing practical and intelligent manipulation closer to everyday applications. The outcomes of this work lay the foundation for future advancements in robotic grasping systems moving closer to a future where robots can autonomously handle complex tasks in unpredictable environments with greater autonomy and precision.

References

[1] Murphy, R. R. (2004). Human-robot interaction in rescue robotics. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34(2), 138-153.

[2] Tenorth, M., & Beetz, M. (2017). Representations for robot knowledge in the KnowRob framework. Artificial Intelligence, 247, 151-169.

[3] Siciliano, B. (2008). Springer Handbook of Robotics. Springer-Verlag google schola, 2, 15-35.

[4] Kazanzides, P., Fichtinger, G., Hager, G. D., Okamura, A. M., Whitcomb, L. L., & Taylor, R. H. (2008). Surgical and interventional robotics-core concepts, technology, and design [tutorial]. IEEE robotics & automation magazine, 15(2), 122-130.

[5] Qin, Y., Chen, R., Zhu, H., Song, M., Xu, J., & Su, H. (2020, May). S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes. In Conference on robot learning (pp. 53-65). PMLR.

[6] Bohg, J., Morales, A., Asfour, T., & Kragic, D. (2013). Data-driven grasp synthesis—a survey. IEEE Transactions on robotics, 30(2), 289-309.

[7] Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., & Quillen, D. (2018). 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.

[8] Hermans, T., Li, F., Rehg, J. M., & Bobick, A. F. (2013, August). Learning stable pushing locations. In 2013 IEEE third joint international conference on development and learning and epigenetic robotics (ICDL) (pp. 1-7). IEEE.

[9] Alliegro, A., Rudorfer, M., Frattin, F., Leonardis, A., & Tommasi, T. (2022). End-to-end learning to grasp via sampling from object point clouds. IEEE Robotics and Automation Letters, 7(4), 9865-9872.

[10] Mahler, J., Liang, J., Niyaz, S., Laskey, M., Doan, R., Liu, X., ... & Goldberg, K. (2017). Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. arXiv preprint arXiv:1703.09312.

[11] Morrison, D., Corke, P., & Leitner, J. (2018). Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. arXiv preprint arXiv:1804.05172.

[12] Brook, P., Ciocarlie, M., & Hsiao, K. (2011, May). Collaborative grasp planning with multiple object representations. In 2011 IEEE international conference on robotics and automation (pp. 2851-2858). IEEE.

[13] Sundermeyer, M., Mousavian, A., Triebel, R., & Fox, D. (2021, May). Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 13438-13444). IEEE.

[14] Wang, C., Fang, H. S., Gou, M., Fang, H., Gao, J., & Lu, C. (2021). Graspness discovery in clutters for fast and accurate grasp detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 15964-15973).

[15] Zhao, B., Zhang, H., Lan, X., Wang, H., Tian, Z., & Zheng, N. (2021, May). Regnet: Region-based grasp network for end-to-end grasp detection in point clouds. In 2021 IEEE international conference on robotics and automation (ICRA) (pp. 13474-13480). IEEE.

[16] Wu, C., Chen, J., Cao, Q., Zhang, J., Tai, Y., Sun, L., & Jia, K. (2020). Grasp proposal networks: An end-to-end solution for visual learning of robotic grasps. Advances in Neural Information Processing Systems, 33, 13174-13184.

[17] Wei, W., Luo, Y., Li, F., Xu, G., Zhong, J., Li, W., & Wang, P. (2021, May). Gpr: Grasp pose refinement network for cluttered scenes. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 4295-4302). IEEE.

[18] Aldaco, J., Armstrong, T., Baruch, R., Bingham, J., Chan, S., Draper, K., Dwibedi, D., Finn, C., Florence, P., Goodrich, S., Gramlich, W., Hage, T., Herzog, A., Hoech, J., Nguyen, T., Storz, I., Tabanpour, B., Takayama, L., Tompson, J., Wahid, A., Wahrburg, T., Xu, S., Yaroshenko, S., Zakka, K., & Zhao, T. (2024). *ALOHA 2: An Enhanced Low-Cost Hardware for Bimanual Teleoperation* [Website]. GitHub Pages. Available at: `https://aloha-2.github.io/`

[19] Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems, 30.

[20] Alliegro, A., Valsesia, D., Fracastoro, G., Magli, E., & Tommasi, T. (2021). Denoise and contrast for category agnostic shape completion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4629-4638).

[21] Alliegro, A., Rudorfer, M., Frattin, F., Leonardis, A., & Tommasi, T. (2022). *L2G* [GitHub repository]. GitHub. Available at: `https://github.com/antoalli/L2G`

[22] Berenson, D., Srinivasa, S. S., & Kuffner, J. J. (2009, October). Addressing pose uncertainty in manipulation planning using task space regions. In 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 1419-1425). IEEE.

[23] Saxena, A., Wong, L. L., & Ng, A. Y. (2008, July). Learning grasp strategies with partial shape information. In AAAI (Vol. 3, No. 2, pp. 1491-1494).

[24] Balasubramanian, R., Xu, L., Brook, P. D., Smith, J. R., & Matsuoka, Y. (2010, May). Human-guided grasp measures improve grasp robustness on physical robot. In 2010 IEEE International Conference on Robotics and Automation (pp. 2294-2301). IEEE.

[25] Fang, H. S., Wang, C., Gou, M., & Lu, C. (2020). Graspnet-1billion: A large-scale benchmark for general object grasping. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11444-11453).

[26] Montemerlo, M., Roy, N., & Thrun, S. (2003, October). Perspectives on standardization in mobile robot programming: The Carnegie Mellon navigation (CARMEN) toolkit. In Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453) (Vol. 3, pp. 2436-2441). IEEE.

[27] Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., ... & Ng, A. Y. (2009, May). ROS: an open-source Robot Operating System. In ICRA workshop on open source software (Vol. 3, No. 3.2, p. 5).

[28] Shehawy, H., Rocco, P., & Zanchettin, A. M. (2021, December). Estimating a garment grasping point for robot. In 2021 20th International Conference on Advanced Robotics (ICAR) (pp. 707-714). IEEE.

[29] Wang, J., & Olson, E. (2016, October). AprilTag 2: Efficient and robust fiducial detection. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 4193-4198). IEEE.

[30] Maitin-Shepard, J., Cusumano-Towner, M., Lei, J., & Abbeel, P. (2010, May). Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In 2010 IEEE International Conference on Robotics and Automation (pp. 2308-2315). IEEE.

[31] Saxena, A., Driemeyer, J., Kearns, J., & Ng, A. (2006). Robotic grasping of novel objects. Advances in neural information processing systems, 19.

[32] Saxena, A., Driemeyer, J., Kearns, J., Osondu, C., & Ng, A. Y. (2008). Learning to grasp novel objects using vision. In Experimental Robotics: The 10th International Symposium on Experimental Robotics (pp. 33-42). Springer Berlin Heidelberg.

[33] Saxena, A., Wong, L., Quigley, M., & Ng, A. Y. (2011). A vision-based system for grasping novel objects in cluttered environments. In Robotics Research: The 13th International Symposium ISRR (pp. 337-348). Springer Berlin Heidelberg.

[34] Saxena, A., Driemeyer, J., & Ng, A. Y. (2009, May). Learning 3-d object orientation from images. In 2009 IEEE International conference on robotics and automation (pp. 794-800). IEEE.

[35] Saxena, A., Driemeyer, J., & Ng, A. Y. (2008). Robotic grasping of novel objects using vision. The International Journal of Robotics Research, 27(2), 157-173.

[36] Schmidt, P., Vahrenkamp, N., Wächter, M., & Asfour, T. (2018, May). Grasping of unknown objects using deep convolutional neural networks based on depth images. In 2018 IEEE international conference on robotics and automation (ICRA) (pp. 6831-6838). IEEE.

[37] Kasper, A., Xue, Z., & Dillmann, R. (2012). The kit object models database: An object model database for object recognition, localization and manipulation in service robotics. The International Journal of Robotics Research, 31(8), 927-934.

[38] Calli, B., Walsman, A., Singh, A., Srinivasa, S., Abbeel, P., & Dollar, A. M. (2015). Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. arXiv preprint arXiv:1502.03143.

[39] Liang, H., Ma, X., Li, S., Görner, M., Tang, S., Fang, B., ... & Zhang, J. (2019, May). Pointnetgpd: Detecting grasp configurations from point sets. In 2019 International Conference on Robotics and Automation (ICRA) (pp. 3629-3635). IEEE.

[40] Dovrat, O., Lang, I., & Avidan, S. (2019). Learning to sample. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2760-2769).

[41] Qin, Y., Chen, R., Zhu, H., Song, M., Xu, J., & Su, H. (2020, May). S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes. In Conference on robot learning (pp. 53-65). PMLR.

[42] Nguyen, V. D. (1988). Constructing force-closure grasps. The International Journal of Robotics Research, 7(3), 3-16.

[43] Yan, X., Hsu, J., Khansari, M., Bai, Y., Pathak, A., Gupta, A., ... & Lee, H. (2018, May). Learning 6-dof grasping interaction via deep geometry-aware 3d representations. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3766-3773). IEEE.