

Dexgrasp

A project of the 2024 Robotics Course of the School of Information Science and Technology (SIST) of
ShanghaiTech University

<https://robotics.shanghaitech.edu.cn/teaching/robotics2024>

Yingdong Shi[†], Heng Tao[†], Youzhuo Wang[†]
ShanghaiTech University

Abstract

This study presents a novel multi-modal approach for precise object manipulation using robotic arms, integrated with advanced computer vision techniques. Utilizing a configuration of four Kinect sensors, our system captures both RGB and depth images surrounding a desktop environment. Employing Llava-Next combined with specific prompts such as "Identify the items above the table in the scene," our model efficiently recognizes and names the objects placed on the table. These identified items are then segmented from the RGB images using the Grounding Dino framework, resulting in precise object masks. These masks are subsequently applied to the depth images to crop the exact object regions, allowing the extraction of accurate point clouds of the tabletop objects from multiple Kinect perspectives. Prior to these procedures, each object had been scanned with a high-precision industrial scanner to create a detailed mesh model. By performing global registration followed by iterative closest point (ICP) adjustments between the object point clouds and mesh data, we accurately determine the objects' poses. These poses are then fed into a trained conditional variational autoencoder (CVAE), which predicts potential grasping poses. These poses serve as inputs to our auto-regressive motion network, Motion-Net, which generates the motion trajectories for the robotic arms to execute precise object grasping. This integrated approach showcases significant improvements in the automation and accuracy of robotic handling in dynamic and cluttered environments.

1 Introduction

The dynamic interaction between robotic systems and complex environments poses significant challenges, especially in the precise manipulation and grasping of objects. Advancements in sensor technology, machine learning, and computer

vision have paved the way for significant improvements in this domain. This paper introduces a sophisticated integration of multi-modal sensing and advanced computational techniques aimed at enhancing the accuracy and adaptability of robotic arms in object manipulation tasks.

The core of our system utilizes an array of four Kinect sensors strategically positioned around a desktop to capture comprehensive RGB and depth images. This setup not only provides a holistic view of the environment but also facilitates the detailed capture of objects placed on the table. The initial step in our process involves object recognition, achieved through Llava-Next, which utilizes natural language prompts to accurately identify and name objects within the captured images. Subsequent segmentation of these objects from the RGB data is executed using the Grounding Dino framework, a technique that effectively isolates the objects and generates corresponding masks. These masks are then applied to the depth images, allowing for the extraction of precise point clouds representing the objects from different angles.

To ensure the accuracy of object representations, we previously scanned each object with a high-precision industrial scanner to create detailed mesh models. These meshes serve as a foundational comparison for the point clouds obtained from the Kinect sensors. Through a series of global registration and iterative closest point (ICP) procedures, we meticulously align and adjust the Kinect-generated point clouds with the mesh models to determine the exact object poses.

Building upon the foundational technologies developed in our prior work, "RealDex: Towards Human-like Grasping for Robotic Dexterous Hand", we employ a trained conditional variational autoencoder (CVAE), referred to as GraspNet, to generate feasible grasping poses. These poses are inputs for our auto-regressive motion network, Motion-Net, which synthesizes the motion trajectories necessary for the robotic arms to perform precise grasping actions. This integration of GraspNet and Motion-Net, derived from our earlier innovations, illustrates a seamless blend of object recognition, pose estimation, and motion planning that significantly enhances the robotic arm's capability to replicate human-like grasping movements in a cluttered and dynamic environment.

In conclusion, this paper details the development and implementation of a robust system that leverages cutting-edge technology in computer vision, artificial intelligence, and robotics to achieve superior object manipulation, demonstrat-

[†]These authors contribute equally to this work.

*

ing potential applications in various industrial, domestic, and commercial settings.

2 State of the Art

2.1 Heng tao

Literature Relevant to This Project

Grasp Pose Detection in Point Clouds

<https://ar5iv.org/abs/1706.09911> This paper discusses various methods for grasp detection directly from sensor data, focusing on kernel representations to encode local geometry of object surfaces. Deep learning-based methods are emphasized for their speed and efficiency in evaluating neural networks for grasp detection, highlighting the computational challenges of kernel density estimators

Deep Robotic Grasping Prediction with Hierarchical RGB-D Fusion

<https://ar5iv.org/pdf/1909.06585v1> This study introduces the U-Grasping Network V3 (UG-Net V3), which employs a hierarchical fusion of RGB and depth data for robotic grasping. It features multimodal fusion, background extraction, and multi-task learning structures for grasp representation. This approach addresses the challenge of grasping with partial observation and without full 3D prior knowledge

Visual Robotic Object Grasping Through Combining RGB-D Data

<https://github.com/atenpas/gpd> Proposes a novel framework for automated robotic grasp by matching captured RGB-D data with 3D meshes. This framework consists of two modules focusing on pre-defining grasping knowledge for object shapes and automated grasp decision-making

Detailed Paper Overview

<https://ar5iv.org/abs/1706.09911> The Grasp Pose Detection in Point Clouds paper presents an insightful exploration into the realm of robotic grasp detection, leveraging kernel density estimators to discern the local geometry of object surfaces for grasp prototyping. The technique's comparison with deep learning-based methods reveals a stark contrast in computational efficiency, marking a significant advancement in the grasp detection domain. This paper is pivotal for your project as it directly addresses the intricacies of grasp pose detection in point clouds, offering a deep dive into the challenges of kernel representations and the superiority of neural network evaluations. It sets a foundational understanding, demonstrating the evolution of grasp detection methodologies and offering a comprehensive problem statement and algorithm for grasp pose detection

Relevant Open Source ROS Package

<https://github.com/atenpas/gpd> GPD (Grasp Pose Detection) is an open-source ROS package specifically designed for detecting 6-DOF grasp poses in point clouds. Key features include its applicability to novel objects without the need for CAD models, effectiveness in dense clutter, and capability to output 6-DOF grasp poses for a 2-finger robot hand. The package follows a two-step process of sampling a large number of grasp candidates and classifying these candidates into

viable grasps or not. It requires PCL 1.9 or newer, Eigen 3.0 or newer, and OpenCV 3.4 or newer for installation and has been tested on Ubuntu 16.04. GPD is particularly suitable for applications where top-down grasps are insufficient, offering a more versatile approach to robotic grasping. The package's architecture allows for significant flexibility in grasp detection, catering to a wide range of robotic applications

2.2 Yingdong Shi

Literature Relevant to This Project

Bilateral Cross-Modal Fusion Network for Robot Grasp Detection

<https://www.mdpi.com/1424-8220/23/6/3340> This paper introduces a tri-stream cross-modal fusion architecture designed to leverage RGB and depth data efficiently for robotic grasp detection. It features a novel modal interaction module (MIM) and channel interaction modules (CIM) to enhance the aggregation of different modal streams. The architecture demonstrated remarkable accuracy in grasp detection across standard datasets and real robot experiments

RGB-D Grasp Detection via Depth Guided Learning with Cross-modal Attention

<https://www.mdpi.com/1424-8220/23/6/3340> This study proposes the DGCAN framework, a depth-guided learning framework that includes a Local Cross-modal Attention (LCA) module for multi-modal fusion. The framework utilizes a 6-dimensional rectangle representation for grasp detection, adding grasp depth as a crucial parameter. It emphasizes asymmetric fusion to refine depth features through cross-modal relation learning

AsymFormer: Asymmetrical Cross-Modal Representation Learning for Mobile Platform Real-Time RGB-D Semantic Segmentation

<https://www.mdpi.com/1424-8220/23/6/3340> AsymFormer is a novel network designed for real-time RGB-D semantic segmentation on mobile platforms. It uses a ConvNext based backbone for RGB feature extraction and a Mix-Transformer based backbone for RGB-D fused features processing, along with Local Attention-Guided Feature Selection (LAFS) and Cross Modal Attention-Guided Feature Correlation Embedding (CMA) modules for improved feature fusion and segmentation performance

Detailed Overview of a Paper

<https://www.mdpi.com/1424-8220/23/6/3340> The Bilateral Cross-Modal Fusion Network for Robot Grasp Detection paper presents an innovative approach to robotic grasp detection by efficiently combining RGB and depth data. Through its tri-stream cross-modal fusion architecture, it facilitates in-depth interaction between RGB and depth information, capitalizing on the strengths of both modalities to enhance detection accuracy. The architecture incorporates MIM for global association information capture between modalities and CIM units for refining the aggregation of modal streams. This approach not only addresses the challenge of multimodal fusion in robotic grasp detection but also showcases superior accuracy in practical applications, making it a valuable reference for your project

Relevant Open Source ROS Package or Software Library

For the practical implementation of robotic grasp detection in your project, considering the advancements and methodologies discussed in these papers, you might want to look into existing ROS packages or software libraries that focus on RGBD data processing and manipulation. While the papers do not specify a ROS package directly related to their proposed methods, the Point Cloud Library (PCL) and OpenCV are extensively used in the field for handling RGBD data and could be instrumental in implementing algorithms based on these studies. Additionally, exploring ROS packages that offer grasp detection functionalities, like the MoveIt Grasp Planning Pipeline, could provide a starting point for integrating these advanced methods into a robotic system.

2.3 Youzhuo Wang

Bilateral Cross-Modal Fusion Network for Robot Grasp Detection

<https://www.mdpi.com/1424-8220/23/6/3340> explores a tri-stream cross-modal fusion architecture designed to efficiently leverage RGB and depth data for robotic grasp detection. The architecture utilizes a novel modal interaction module (MIM) and channel interaction modules (CIM) to enhance multiscale information aggregation, achieving high accuracy in standard datasets and real robot experiments

Deep Robotic Grasping Prediction with Hierarchical RGB-D Fusion

<https://ar5iv.org/pdf/1909.06585> presents a method employing a hierarchical encoder-decoder network for robotic grasping prediction based on RGB-D data. It introduces a unique approach to dataset generation and preprocessing, using the Cornell grasp detection dataset and the YCB Object and Model Set, aiming for improved grasping performance through depth image inpainting and multimodal fusion

RGB-D Grasp Detection via Depth Guided Learning with Cross-modal Attention

<https://ar5iv.org/abs/2302.14264> presents a novel approach to improve robotic grasp detection by leveraging RGB-D data. The proposed Depth Guided Cross-modal Attention Network (DGCAN) uses a 6-dimensional rectangle representation to include grasp depth, enhancing feature learning for more accurate results. It introduces a Local Cross-modal Attention module to refine depth features through cross-modal relation learning and concatenate them to RGB features for fusion. The method has shown superior performance in both simulation and physical evaluations

Detailed Overview of a Paper

<https://www.mdpi.com/1424-8220/23/6/3340> The paper "Bilateral Cross-Modal Fusion Network for Robot Grasp Detection" offers an in-depth analysis and solution to the challenge of accurately determining the position and pose of a target in robotic grasp detection by leveraging both RGB and depth information. The proposed architecture significantly contributes to the field by enabling efficient interaction between RGB and depth data, enhancing the aggregation of multiscale information through its tri-stream cross-modal fusion design. The modal interaction module (MIM) within the architecture

is pivotal for adaptively capturing cross-modal feature information, while channel interaction modules (CIM) further refine the aggregation process. This structure has been validated through experiments on standard public datasets, such as the Cornell and Jacquard datasets, where it demonstrated exceptional image-wise and object-wise detection accuracy. Furthermore, the practical application of this architecture in physical experiments with the 6-DoF Elite robot showcased a high success rate, underscoring the method's real-world applicability and effectiveness

Relevant Open Source ROS Package or Software Library

<https://github.com/atenpas/gpd> The Grasp Pose Detection (GPD) package is a robust tool designed for detecting 6-DOF grasp poses in 3D point clouds, suitable for a two-finger robot hand. It's capable of working with novel objects without pre-existing CAD models, operates effectively in dense clutter, and outputs 6-DOF grasp poses, enabling more sophisticated grasping strategies beyond simple top-down approaches. GPD performs by first sampling a large number of grasp candidates, then classifying these candidates to identify viable grasps. This package is essential for projects requiring advanced grasp detection capabilities in complex environments

3 a paper relevant to my problem

3.1 UnidexGrasp

The paper titled "UniDexGrasp: Universal Robotic Dexterous Grasping via Learning Diverse Proposal Generation and Goal-Conditioned Policy" primarily addresses the challenge of universal robotic dexterous grasping using a novel two-stage learning approach. Here's a summary of the key aspects presented:

Abstract

The paper presents a system designed to perform dexterous grasping in a table-top setting, with the objective to handle objects in diverse and high-quality manners across a wide range of categories, including those unseen during training. The approach consists of two main stages: (1) grasp proposal generation based on point cloud data, and (2) goal-conditioned grasp execution.

Introduction

The need for dexterous robotic grasping capabilities far surpasses what current technologies with parallel grippers can offer, especially regarding handling complex and functional object manipulations. This paper introduces an innovative solution that significantly improves upon the limitations of existing grasping approaches by using a more versatile and capable system.

System Overview

Grasp Proposal Generation: A novel probabilistic model predicts diverse grasp poses based on the observed point cloud. This model separates the rotation, translation, and articulation of grasps to enhance the quality and diversity of the proposals. **Goal-Conditioned Grasp Execution:** Replaces traditional motion planning with a goal-conditioned policy to manage the

complexity of dexterous grasping, aiming to achieve a high generalization capability.

Algorithm Details

Machine Learning Model: The approach utilizes deep learning to predict and execute grasp maneuvers. Techniques such as convolutional neural networks (CNNs) are employed for analyzing point cloud data, while recurrent neural networks (RNNs) manage sequential decision-making tasks. **Data Handling and Training:** Trained on a synthesized dataset comprising over one million grasp instances, the model learns to generalize across different objects by understanding the varying conditions of grasping.

Key Innovations

Probabilistic Grasp Pose Modeling: The model separates the analysis of rotation from translation and articulation, allowing for more sophisticated handling of the intrinsic high-dimensionality of dexterous manipulations. **Goal-Conditioned Policy:** Enables the system to execute grasps based on predetermined goal poses, enhancing adaptability and effectiveness in real-world applications.

Results

The system demonstrates an average success rate of over 60% across thousands of object instances, significantly outperforming traditional methods. This performance showcases the system’s capability to adapt and generalize well beyond the training scenarios.

Conclusion

“UniDexGrasp” marks a significant step forward in robotic dexterous grasping technology, showing promising results in both academic and potentially commercial robotics applications. The methodology and results pave the way for future research and development in robotic manipulation.

This paper represents a comprehensive effort to tackle one of the more challenging aspects of robotics and automation, with potential implications for industries requiring high precision and variability in object handling.

4 System Description

4.1 System and Algorithm Details

Multi-Modal Sensing Setup

The system employs a configuration of four Kinect sensors surrounding a desktop area. This setup captures both RGB and depth information from multiple angles, ensuring comprehensive visibility and data collection of the tabletop environment. The Kinect sensors are calibrated to synchronize their feeds, allowing for simultaneous capturing of the scene from different viewpoints.



Figure 1: hardware set up.

Object Recognition and Segmentation

Using the software Llava-Next, the system processes natural language prompts, such as “Identify the items above the table in the scene.” This AI-driven approach leverages deep learning models trained on vast datasets to recognize and label objects within the RGB images accurately. Once identified, the Grounding Dino framework is utilized to segment these objects from the RGB images. Grounding Dino leverages grounding techniques to pinpoint and create masks around the named entities, isolating them from the rest of the image.

Depth Image Processing

The masks obtained from the RGB segmentation are applied to the corresponding areas in the depth images. This step isolates the depth information of the objects alone, allowing for the extraction of detailed 3D point clouds. These point clouds represent the exact shape and size of the objects as captured by the Kinect sensors.

Point Cloud Registration and Mesh Comparison

Each object on the table is previously scanned using a high-precision industrial scanner to create a mesh model. The system performs global registration followed by Iterative Closest Point (ICP) algorithms to align and adjust the Kinect-generated point clouds with these pre-scanned mesh models. This step is crucial for accurate pose estimation and ensures the physical dimensions and orientations of the objects are precisely known.

Pose Prediction and Grasping

The aligned point clouds and their accurate poses feed into a Conditional Variational Autoencoder (CVAE), termed Grasp-Net, which is trained to predict potential grasping poses. This prediction model is based on learning from numerous examples of effective grasps, tailored to the specific shapes and sizes of objects.

Motion Planning

The predicted poses are then input to an auto-regressive motion network, Motion-Net, which generates the detailed motion trajectories required for a robotic arm to reach and grasp the objects. This network is trained to consider the dynamics and kinematics of the robotic arm, ensuring smooth and feasible movements that can be executed in real time.

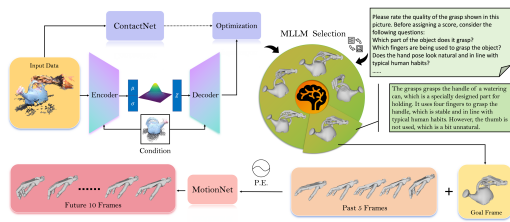


Figure 2: The architecture of our grasping motion generation framework. When observing the point cloud of the object, a cVAE-based generation module is used to generate multiple grasping pose candidates. Then, a MLLM selection module is utilized to select the most reasonable and human-like pose. Finally, based on the goal grasp pose, the MotionNet synthesizes the motion sequence for robot execution.

ROS Integration

In our project, we use ROS (Robot Operating System) packages specifically tailored for the Shadow Hand robotic system to manage its complex movements and sensor integration. Additionally, we integrate Open3D for efficient 3D data processing, which is crucial for tasks like object recognition and pose estimation. Our custom ROS nodes facilitate the conversion of sensor data into usable formats, handle detailed data processing, and execute movement commands based on the data processed. Integrating these technologies posed challenges in ensuring real-time performance and robust synchronization, which we addressed through system optimization and communication enhancements. This setup forms a sophisticated platform capable of performing advanced robotic manipulation tasks.

4.2 Overcome Challenges

Complex Object Recognition in Cluttered Environments

The integration of Llava-Next with specific prompts helps overcome the challenge of recognizing and labeling objects accurately in cluttered tabletop scenarios. This is particularly difficult where multiple objects might overlap or obscure each other in typical RGB and depth images.

Precise Object Segmentation from Multi-Modal Data

The application of Grounding Dino to segment objects from RGB data based on their names and then applying these segments as masks on depth data required innovative approaches to ensure precision. This was crucial for isolating the exact object shapes needed for accurate point cloud generation.

Alignment of Point Clouds with High-Precision Mesh Models

The variability in data capture quality between consumer-grade Kinect sensors and industrial scanners posed a significant challenge. Employing robust registration and ICP algorithms helped refine the alignment process, ensuring that the point clouds closely matched the high-fidelity mesh models.

Adaptive and Predictive Grasping

Developing GraspNet to predict grasping poses involved overcoming the challenge of variable object shapes and sizes, necessitating a model that could adaptively learn from diverse training examples to generalize well across unseen objects.

Real-Time and Dynamic Motion Planning

Implementing Motion-Net required solving the problem of generating real-time, dynamic motion trajectories that are both feasible and efficient for robotic arms. This involved intricate modeling of the arm’s mechanics and the physics of movement to ensure the trajectories are safe and effective for execution.

4.3 System Evaluation

My test involves inputting an object model without human intervention and performing pose estimation on point clouds captured by either four merged Kinects or a single Kinect. If the object model aligns well with the object in the point cloud, then the algorithm test is considered successful. We have achieved good results with many objects we tested. Another test involves specifying the pose of an object and having a robotic hand grasp the object in the point cloud, which also yielded good results.



Figure 3: only one camera kinect point cloud register spirit.

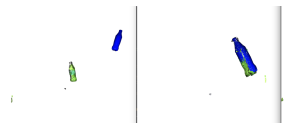


Figure 4: four camera kinect point cloud register spirit

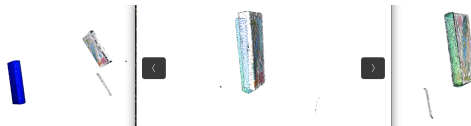


Figure 5: four camera kinect point cloud register bowling game.

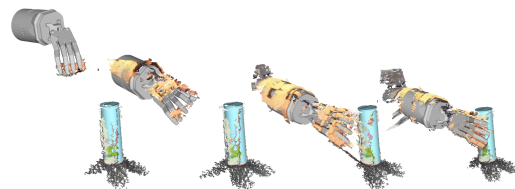


Figure 6: shadow hand grasp object in real world.

4.4 How to

The code to obtain a clean point cloud of just the object can be accessed in

```
code/get_rgb_mask/segmentation/  
grounded_sam.ipynb.
```

Using the clean object point cloud and the object's mesh, code to achieve accurate object pose through global registration and ICP is available in

```
code/grounded_sam.ipynb.
```

The model that uses the object's point cloud to determine the robotic hand's pose and motion for grasping can be trained using the script

```
code/dexgrasp_generation/train.sh
```

to obtain the grasping and motion model

4.5 Summary

In our study, we developed a novel method for precise robotic object manipulation using a multi-sensor setup and advanced image processing techniques. We utilized four Kinect sensors positioned around a desktop to capture both RGB and depth images. Using the Llava-Next software with prompts such as "Identify the items above the table in the scene," we identified objects on the table and segmented them from the RGB images using a tool named Grounding Dino. This tool created masks for the objects, which we then applied to the depth images to extract accurate object regions.

We processed images from all four Kinect sensors in this manner to produce precise point clouds of the tabletop objects. Prior to this, each object had been scanned with a high-precision industrial scanner to create a detailed mesh model. We performed global registration and iterative closest point (ICP) adjustments on the point clouds and mesh models to determine the accurate poses of the objects under the multiple Kinect setup.

These poses were fed into a trained Conditional Variational Autoencoder (CVAE), which generated potential grasping poses for the objects. These grasping poses were then input into an auto-regressive motion network, Motion-Net, to create the necessary motion trajectories for a robotic hand to execute precise grasps. This integrated approach allows the robotic hand to effectively grasp identified objects, demonstrating significant advancements in automated object handling.

4.6 Future Work

To further enhance the system, several avenues could be pursued

Sensor Fusion Enhancements

Integrating additional types of sensors, such as LiDAR or more advanced depth cameras, could improve the quality and reliability of the captured data, leading to even more precise object recognition and segmentation.

Algorithm Optimization

Continuous improvements in the algorithms for segmentation, point cloud registration, and pose estimation could increase the accuracy and speed of the system. Leveraging recent advancements in AI and machine learning could provide new ways to optimize these processes.

Grasp and Motion Planning

Further development of GraspNet and Motion-Net to include feedback loops could allow the system to adapt to failed attempts in real-time, enhancing its learning and adjustment capabilities during operation.

Extended Object Library

Expanding the training datasets with a broader array of objects and grasp scenarios could significantly enhance the generalizability and robustness of the grasping predictions, making the system more versatile across different settings.

References