Force Closure-Aware Grasp Detection with PointNet++ Feature Extraction

Zimo Wen
2581235653@sjtu.edu.cn

ABSTRACT

Dexterous grasp detection is a fundamental task in robotic learning, crucial for achieving stable object grasping. The key challenge lies in extracting effective information for prediction. To fully address this issue, we propose a novel force-based grasp detection model, DexGraspDetector, which employs PointNet++ for point cloud feature extraction and incorporates force-closure analysis. Utilizing an attention mechanism, the model dynamically balances the relationships among point cloud, pose, and joint features. Experimental results demonstrate that our model excels in grasp prediction accuracy, achieving a precision of 93.2%, highlighting the effectiveness of combining deep learning algorithms with physical stability analysis.

KEYWORDS

Dexterous Grasp Detection, Force-Closure, Robotic Manipulation,Attention Mechanism

1 INTRODUCTION

Grasp detection is a fundamental task in robotic learning, aiming to determine whether an object can be stably grasped based on inputs such as 3D point clouds and joint poses. While significant progress has been made in grasp detection for simple grippers, dexterous grasping presents unique challenges due to the high degrees of freedom, complex spatial relations, and intricate mechanical structures of multi-fingered robotic hands. For instance, tasks involving high-DoF robotic hands, such as the Shadow Hand, require the model to handle both the spatial relationships between the hand and the object and the mechanical interactions between fingers,as shown in figure 6 . This increases the complexity of grasp prediction and makes it more sensitive to diverse object shapes and complex scenarios.

To address these challenges, we propose a novel framework that combines feature extraction with physical stability analysis for dexterous grasp detection. Our approach first computes the force-closure metrics of grasps based on physical principles. After calculating the force-closure metrics, we use a feed-forward network (FFN) and PointNet++ to process joint information and point clouds, respectively. Finally, we employ an attention mechanism to dynamically balance the relationships between point cloud features, poses, and joint features based on the computed force-closure metrics. This combination allows the model to consider both the geometric structure of the object and the physical properties of grasping.

The primary problem we address is how to effectively integrate multimodal data—including point clouds, grasp poses, and joint configurations—while accounting for the physical stability required for dexterous manipulation. Existing methods often struggle to balance these aspects, leading to suboptimal performance in complex grasping scenarios. Our framework overcomes this limitation by introducing an attention mechanism that dynamically balances the contributions of geometric, pose, and joint features based on force-closure metrics, ensuring that the model focuses on the most relevant features for grasp success in each scenario.

Experimental results demonstrate that our method achieves a grasp prediction accuracy of 93.2% on multiple benchmark datasets. This not only validates the effectiveness of our approach but also highlights the immense potential of combining geometric learning with physical stability analysis, offering a new solution to the challenges of dexterous grasp detection.

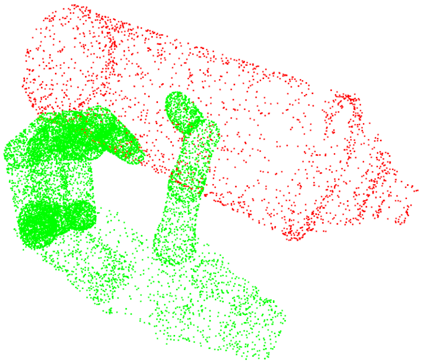

Figure 1: Dexterous Hand Grasping Diagram

2 RELATED WORK

2.1 Geometric Feature Extraction and Deep Learning

Geometric feature extraction is fundamental to grasp detection. Traditional methods often rely on geometric modeling or predefined features to estimate grasp candidates. PointNet [3] and PointNet++ [4] pioneered direct processing of point cloud data, laying the groundwork for geometry-based grasp detection. Building on these techniques,significantly improving performance.

2.2 Physical Stability Analysis and Multimodal Fusion

Successful grasp detection requires not only geometric alignment but also the evaluation of physical stability. Bohg et al. [1] provided a comprehensive survey of force-closure and form-closure theories, which form the theoretical basis for grasp stability. Recent studies have incorporated physical stability into deep learning frameworks. For example, DexNet [?] integrated point cloud features with force-closure metrics to achieve robust grasp detection. However, the dynamic balancing of multimodal features remains a challenge in complex scenarios.

2.3 Large-Scale Simulation Data and Benchmarks

The introduction of large-scale simulation datasets has significantly advanced research in dexterous grasp detection. GraspNet-1Billion [2] is the largest grasp dataset to date, containing over one billion simulated grasp samples across diverse objects and scenes. It leverages simulation environments to generate diverse grasp strategies and filters stable grasps using force-closure metrics. GraspNet-1Billion provides a unified benchmark for training and evaluation, enhancing model generalization in complex scenarios. Additionally, simulation tools like PyBullet and Mujoco offer flexible platforms for dynamic simulation and evaluation of high-DoF robotic hands.

In summary, the grasp detection algorithms for robotic hands have been continuously advancing and are widely applied in areas such as grasp strategy evaluation and dynamic simulation.

3 METHODOLOGY

In this section, we introduce our grasp model framework, which leverages force-closure, an attention mechanism, and PointNet++. First, we define the problem addressed by the model in Section 3.1. Next, we describe the overall architecture of the model in Section 3.2. Then, we provide a detailed analysis of the main components of the model in Sections 3.3 and 3.4.

3.1 Problem Definition

In this work, we define the task of dexterous grasp detection for high-DoF robotic hands as a multimodal prediction problem. Consider a dataset consisting of multimodal information related to grasp attempts, sampled from N different objects and scenarios. Each data sample includes the following components: a point cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$, representing the geometry of the object; a pose matrix $\mathbf{T} \in \mathbb{R}^{4\times 4}$, describing the spatial position and orientation of the robotic hand; a joint configuration $\mathbf{J} \in \mathbb{R}^{d_J}$, representing the robot's joint states.

The goal is to build a model that takes the multimodal information as input and predicts the success probability of the grasp. Mathematically, the task can be formalized as finding a function f_θ, parameterized by θ, such that

$$f_\theta : (\mathbf{P}, \mathbf{T}, \mathbf{J}) \to p,$$

where $p \in [0, 1]$ represents the probability of a successful grasp.

During training, the model is optimized to minimize a binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right],$$

where B is the batch size, $y_i \in \{0, 1\}$ is the ground truth label indicating grasp success or failure, and $p_i = f_\theta(\mathbf{P}_i, \mathbf{T}_i, \mathbf{J}_i$ is the predicted probability for the i-th sample.

The trained model leverages relationships between multimodal inputs, including geometric structure, hand pose, and physical interactions, to infer grasp feasibility and stability. The objective is to maximize the accuracy and robustness of grasp predictions across diverse objects and scenarios.

3.2 Overall Architecture

Figure 3 illustrates the overall architecture of the model, which comprises force-closure analysis and geometric feature extraction. The force-closure module takes joint and pose data as inputs to compute contact points and force-closure scores, which represent the physical stability of a grasp. While the geometric feature extraction module processes multimodal inputs, including joint, pose, and point cloud data. The joint and pose features are extracted using feed-forward neural (FFN) layers, while the PointNet++ backbone extracts global and local geometric features from the point cloud. These features, weighted dynamically by the attention mechanism using the force-closure score, are fused and passed through a sigmoid function to output the binary prediction of grasp feasibility.

In the following subsections, we will delve into the various components of the model and its overall architecture.

3.3 Force-Closure Analysis

Force closure is a key concept in robotic grasping that quantifies the stability of a grasp by ensuring that contact forces and torques can constrain the object in 3D space. The computation begins with the extraction of **contact points** (c_i) and **contact normals** (n_i), derived from the robot's pose, joint states, and object geometry. These contact normals are further processed to compute two orthogonal **tangential vectors** (t_{i1}, t_{i2}), forming the basis for the friction cone:

$$t_{i1} = \frac{n_i \times a}{\|n_i \times a\|}, \quad t_{i2} = \frac{n_i \times t_{i1}}{\|n_i \times t_{i1}\|}$$

where a is an arbitrary vector not parallel to n_i.

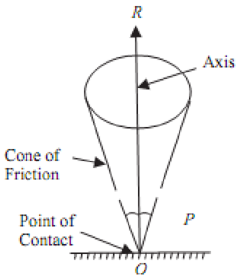

Figure 2: Cone of Friction

Using the tangential vectors and normal vectors, D friction directions (f_d) are sampled within the friction cone:

$$f_d = t_{i1}\cos\phi + t_{i2}\sin\phi + n_i\cos\theta, \quad \theta = \arctan(\mu), \ \phi \in [0, 2\pi]$$

where μ is the friction coefficient, ϕ is the angle within the cone, and θ defines the cone's opening angle.

To evaluate the force closure property, the covariance matrix C is constructed from all sampled friction directions:

$$C = \frac{1}{M \cdot D} F^T F$$

where $F \in \mathbb{R}^{M \cdot D \times 3}$ concatenates friction directions across M contact points. The determinant of C, $\det(C)$, is used as the force closure

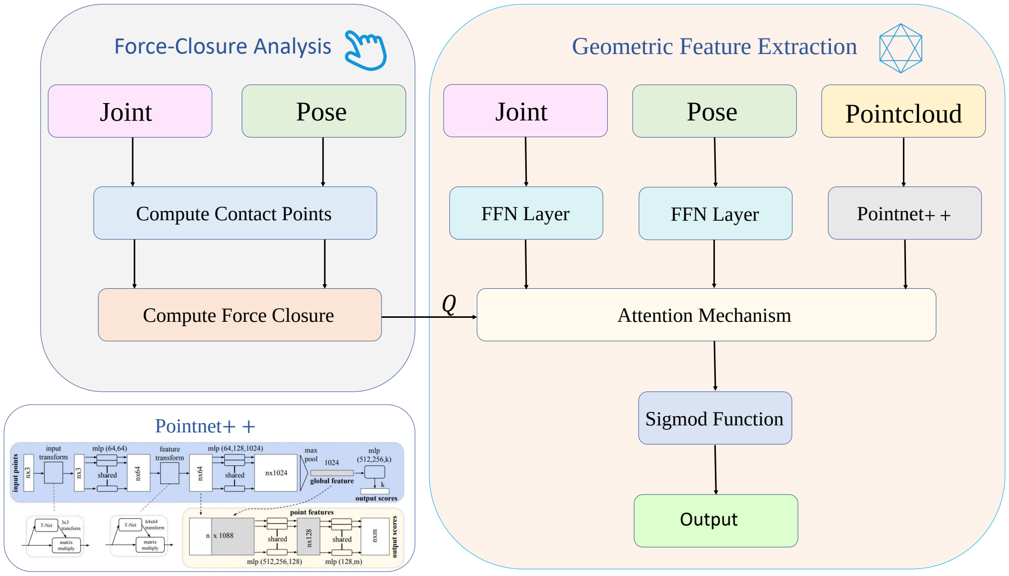

Figure 3: Overall architecture of the model

score:

$$\text{Force Closure Score} = \text{ReLU}(\det(C))$$

A larger determinant indicates a uniform and independent distribution of friction forces, signifying strong force closure and grasp stability. Conversely, $\det(C) = 0$ indicates linear dependence among forces, leading to an unstable grasp.

3.4 Geometric Feature Extraction

The geometric feature extraction module in our framework plays a critical role in analyzing the spatial structure of objects for grasp evaluation. This process begins with the input 3D point cloud, which represents the object's geometry, and extracts both local and global features through hierarchical processing using **PointNet++**. Additionally, the extracted geometric features are dynamically integrated with pose and joint features to ensure a comprehensive understanding of the grasp scenario.

PointNet++ leverages a hierarchical approach to progressively capture local and global geometric features. The first step is arthest point sampling (FPS), which ensures uniform coverage of the input point cloud by selecting a fixed number of representative points. For each sampled point $\mathbf{p}'_j$, its neighborhood is determined using a query ball search, where all points within a radius r are selected:

$$\mathcal{N}(\mathbf{p}'_j) = \{\mathbf{p}_i \in \mathbf{P} \mid \|\mathbf{p}_i - \mathbf{p}'_j\| \le r\}.$$

These local neighborhoods are centered relative to the sampled point:

$$\mathbf{p}_{\text{centered}} = \mathbf{p}_i - \mathbf{p}'_j.$$

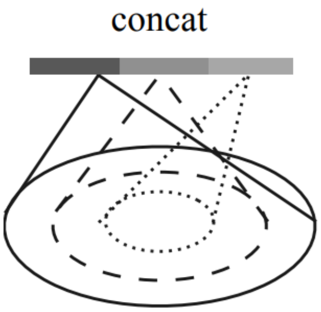

Figure 4: Multi-scale grouping

Once the neighborhood is defined, the local features are learned using shared multi-layer perceptrons (MLPs). For each neighborhood $\mathcal{N}(\mathbf{p}'_j)$, an MLP transforms the input features into higher-dimensional representations:

$$\mathbf{f}_j = \text{MLP}(\mathcal{N}(\mathbf{p}'_j)),$$

followed by a max pooling operation that aggregates the most significant features in the neighborhood:

$$\mathbf{f}'_j = \max(\mathbf{f}_j).$$

This process is repeated across multiple hierarchical layers of set abstraction (SA). Each layer progressively reduces the number of points while increasing the receptive field, culminating in a global feature vector that captures the entire geometry of the object.

In addition to the geometric features, our framework incorporates pose and joint configuration information through separate fully connected layers.These features are combined with the geometric features extracted by PointNet++ to form a comprehensive representation.

To dynamically adjust the importance of each feature, we utilize a force-closure score computed from the contact points and normals of the object. This score serves as the input to an attention mechanism, which assigns weights to the geometric (w_g), pose (w_p), and joint (w_j) features.The final feature vector is computed as a weighted combination:

$$\mathbf{f}_{\text{final}} = w_g \cdot \mathbf{f}_{\text{geom}} + w_p \cdot \mathbf{f}_{\text{pose}} + w_j \cdot \mathbf{f}_{\text{joint}}.$$

This combined feature vector $\mathbf{f}_{\text{final}}$ is passed through fully connected layers and a sigmoid activation function to predict the grasp feasibility:

$$p = \sigma(\text{FC}(\mathbf{f}_{\text{final}})),$$

where y is the binary output indicating whether the grasp is feasible.

By integrating geometric, pose, and joint features dynamically through the attention mechanism, the framework ensures robust and physically-informed grasp evaluations.

4 EXPERIMENTS AND ANALYSIS

4.1 Experimental Setup

In this experiment, we trained and tested the model using the following parameters: the learning rate is set to 0.001, the friction coefficient is 0.5, the number of friction directions is 8, the point cloud feature dimension is 256, the linear transform dimension for pose and joint features is 64, and the combined feature dimension is 256. The batch size is 32, and the model is trained for 30 epochs on an NVIDIA A6000 GPU running Ubuntu 22.04.

4.2 Main Results

Since task-specific models for grasp feasibility prediction are scarce, we conducted ablation studies by removing specific components from our model to validate its effectiveness. We compared our full model, which uses an attention mechanism to dynamically integrate force-closure scores, geometric features, pose features, and joint configurations, with the following variants:

- **Baseline 1: No Force Closure** — The force-closure score was removed from the model to assess its importance in grasp evaluation.
- **Baseline 2: Direct Force Closure Integration** — The force-closure score was directly concatenated with the other features without using the attention mechanism.
- **Baseline 4: Without PointNet++** — The PointNet++ module for geometric feature extraction was replaced with a simple average pooling of the raw point cloud data.
- **Baseline 5: Without Joint and Pose Features** — The pose and joint configuration features were entirely removed, testing the model's performance when only relying on geometric and force-closure features.

The results in Table 1 demonstrate the importance of each component in the proposed model. Our complete model achieves the highest accuracy of **93.2%**. This validates the effectiveness of the attention mechanism in leveraging multimodal inputs for robust grasp evaluation.

Table 1: Main results

Model Variant	Accuracy (%)
Baseline 1: No Force Closure	90.6
Baseline 2: Direct Force Closure Integration	92.4
Our Model: With Attention Mechanism	**93.2**
Baseline 4: Without PointNet++	89.6
Baseline 5: Without Joint and Pose Features	66.1

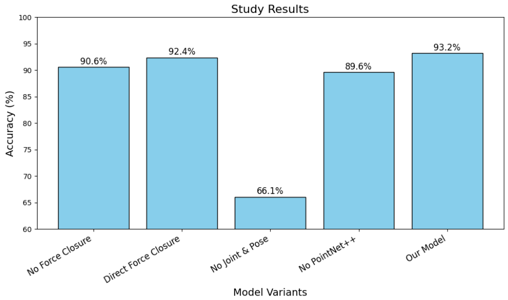

Figure 5: Main results

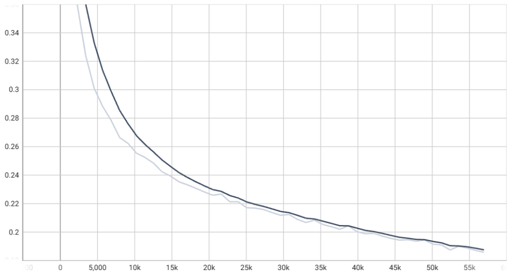

Figure 6: Training Loss Curve

Removing the force-closure score (Baseline 1) reduces the accuracy to **90.6%**, showing its significant role in assessing physical stability. Directly integrating the force-closure score without attention (Baseline 2) improves the accuracy to **92.4%**, but it remains lower than the full model, highlighting the advantage of dynamic feature weighting.

Finally, removing joint and pose features (Baseline 5) causes the largest accuracy drop to **66.1%**, demonstrating that these features are indispensable for understanding the gripper-object relationship and achieving accurate grasp evaluation.

REFERENCES

[1] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. 2013. Data-driven grasp synthesis—a survey. *IEEE Transactions on robotics* 30, 2 (2013), 289–309.

[2] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. 2020. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11444–11453.

[3] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 652–660.

[4] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems* 30 (2017).