Neural Collision Detection for Constrained Grasp Pose Optimization in Cluttered Environments

Longyuan Lin*, Weiwei Zhu*, Yixin Zhuang†, Qinghai Zheng and Yuanlong Yu

***Abstract*— Robust robotic grasping in cluttered environments presents a significant challenge, as existing methods often neglect the complex interactions between the gripper, objects, and obstacles, leading to collisions and grasping failures. To address this, we propose a framework that integrates collision avoidance as a core constraint within the grasp pose optimization process. Central to this framework is a Neural Collision Detection (NCD) network that takes scene configurations and grasp poses as inputs, producing a collision score that approximates traditional collision detection functions. The NCD network provides critical feedback for refining grasp predictions and demonstrates strong generalization across diverse environments, facilitating efficient collision detection and constrained grasp pose optimization. Additionally, we incorporate frictional force closure, geometric symmetry, and surface alignment as regularization terms within the optimization function, enhancing the physical stability and geometric plausibility of the generated grasps. Extensive experiments conducted in real-world environments show a significant improvement in grasp success rates, with robust generalization to previously unseen objects and scenarios. These results validate the efficacy of our framework, highlighting its potential for enabling reliable robotic manipulation in complex and cluttered environments.**

I. Introduction

Robotic grasping in cluttered environments is a critical aspect of modern robotics, with applications ranging from industrial automation and assistive technologies to autonomous systems. To effectively handle objects in these complex scenarios, precise grasp pose optimization is essential. This task is challenging due to the intricate interactions between the robot's gripper, the object, and surrounding obstacles. Even minor inaccuracies in grasp pose predictions can lead to failed grasps, object damage, or potential harm to the robot. Therefore, reliable collision detection is crucial for anticipating and avoiding these risks, ensuring the safety and success of grasping tasks in real-world environments.

Recent advancements in collision-aware grasp pose optimization have made significant progress, yet fully integrating collision detection into the grasping pipeline remains a challenge. Some approaches, such as Pardi et al. [1], address collisions only after grasping, while others, like Lou et al. [2], aim to avoid collisions but are computationally expensive. Methods designed for specific object types, such as Zhang et al. [3] for cables, often lack generalizability. More recent techniques integrate collision avoidance during grasp planning, such as Viturino and Conceicao [4], which relies on point cloud-based distance queries but suffers from high computational costs, and Cai et al. [5], which requires detailed volumetric data. These limitations highlight the need for a more efficient, generalizable approach.

To address these limitations, we propose a novel framework that integrates collision avoidance as a central constraint in the grasp pose optimization process. This framework ensures that potential collisions are proactively accounted for, rather than addressed reactively. At its core is a Neural Collision Detection network, which takes both the scene configuration and the proposed grasp pose as inputs to compute a collision score. In particular, NCD approximates the collision detection function, which is an implicit function. Thus, we model it as neural implicit representations [6]–[8]. This network assesses the likelihood of collisions between the gripper and surrounding objects, providing real-time feedback that allows for adjustments to the grasp pose. By incorporating collision detection early in the optimization process, our approach avoids costly post-collision corrections and enhances grasp reliability.

In addition to collision-aware optimization, we introduce several regularization terms into the optimization objective: frictional force closure, geometric symmetry, and surface alignment. Frictional force closure ensures that the contact forces between the gripper and the object remain within friction cones, promoting grasp stability. Geometric symmetry and surface alignment enhance the gripper's positioning by improving its alignment with the object's centroid and ensuring close proximity to the object's surface.

The integration of collision-aware optimization with stability-enhancing regularization significantly reduces the likelihood of grasp failures caused by collisions, misalignment, or instability, providing a comprehensive solution for grasp pose optimization.

The key contributions of this work are as follows:

- We propose a novel framework that integrates collision avoidance directly into the grasp pose optimization process, ensuring that collisions are proactively accounted for throughout the optimization pipeline.
- We introduce a Neural Collision Detection (NCD) network that approximates collision functions, providing real-time feedback to enhance grasp predictions.
- We include frictional force closure, geometric symmetry, and surface alignment as regularization terms to ensure that the optimized grasps are physically stable and geometrically plausible.

*The first two authors contributed equally to this work.

†Corresponding author.

Longyuan Lin, Weiwei Zhu, Yixin Zhuang, Qinghai Zheng and Yuanlong Yu are with the College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China. This work is supported by the Natural Science Foundation of Fujian Province (2025J01539, 2023J05025), the National Natural Science Foundation of China (U21A20471, 62306074).

II. Related Work

A. Grasp Pose Detection

Recent advancements in robotic grasp pose detection have greatly improved the ability of robotic systems to perform multi-degree-of-freedom (6-DOF) grasps [9]–[12]. Early methods, such as Mahler et al. [13], focused on planar grasp detection, where 2D grasp candidates were evaluated using neural networks. However, these approaches had limitations in capturing the full spatial dynamics of real-world grasping, leading to the development of more advanced 6-DOF techniques.

Methods like GPD [14] and PointNetGPD [15] adopted a sampling-based framework to generate multiple grasp candidates from point cloud data. S4G [16] eliminated the need for exhaustive candidate sampling by directly regressing 6-DOF grasp poses using hierarchical features from PointNet. VGN [17] advanced this further by utilizing Truncated Signed Distance Functions (TSDF) and 3D convolutional neural networks (CNNs) for direct prediction of grasp quality and gripper orientation.

The introduction of large-scale datasets and advanced architectures marked a significant step forward. GraspNet-1Billion [18] provided a large-scale dataset and cascade network architecture, improving the robustness and precision of grasp pose prediction. GSNet [19] introduced a geometry-based quality metric, "graspness", to improve computational efficiency by eliminating low-likelihood grasp points. Ma et al. [20] introduced a hybrid data augmentation strategy and multi-scale cylindrical grouping module to enhance performance for small-scale objects.

Recent works continue to address key challenges in robotic grasp detection. Vergnet improves detection in low-light conditions [21], while knowledge distillation enables smaller, faster models [22]. Attention-augmented networks like AAGDN [23] enhance performance, and increasingly large grasp datasets [24] facilitate further research. Language-driven grasp detection is explored using the conditional consistency model [25], while other efforts focus on generalizing 6-DOF grasps with domain priors [26] and improving grasp strategies in cluttered 7-DOF environments [27].

These contributions trace the evolution from planar grasp detection to advanced 6-DOF techniques, incorporating recent advances in efficiency, attention mechanisms, large-scale datasets, language-driven grasping, and clutter handling, which our approach extends with real-time collision avoidance for constrained grasp pose optimization.

B. Grasp Collision Detection

Collision-aware grasping is a critical challenge in robotics, particularly in constrained or cluttered environments. Numerous methods have been proposed, yet many struggle with real-time performance, generalization, or computational efficiency. For instance, Pardi et al. [1] introduced a framework for post-grasp collision-aware manipulation, handling collisions only after the grasp is executed. While effective in certain scenarios, this approach misses the opportunity to preempt collisions during the grasping phase. In contrast, Lou et al. [2] proposed a target-driven method using the CARP, which estimates the likelihood of a candidate grasp being collision-free through self-supervised simulation training. However, its heavy reliance on prior environmental knowledge and extensive simulation data can limit its adaptability in highly dynamic or unstructured scenarios. Similarly, Zhang et al. [3] focused on collision-aware grasping for flexible objects like cables; though well-suited for such cases, its specialization restricts broader applicability across different geometries.

Recent advancements have aimed to integrate collision avoidance directly into the grasping process. Viturino and Conceicao [4] developed a selective 6D grasping method using point clouds and RGB+D images to detect collisions. However, their approach requires frequent distance queries between the gripper and the environment, leading to significant computational overhead when evaluating numerous grasp candidates. Although simplifying object shapes can reduce computation, it does so at the cost of accuracy and robustness in complex scenes. Likewise, Cai et al. [5] introduced a volumetric-based contact point detection method to mitigate collision risks for 7-DOF robotic arms. This method effectively minimizes collisions but depends on highly detailed TSDF data, which may not be available for irregularly shaped or dynamic objects.

These approaches either fail to provide real-time collision avoidance or have limited generalization across different objects and scenes. Our method integrates collision avoidance directly into the grasp pose optimization process using implicit neural representations to approximate the collision function. This allows for efficient and collision-aware grasp planning without the need for precise 3D models, enhancing reliability, generality, and efficiency across a variety of object types and environments.

III. Method

A. Preliminaries

Let $P = \{p_1, p_2, \ldots, p_n\}$ be the set of 3D points representing the environment, the goal is to find an optimal grasp pose G that results in a stable grasp. A grasp pose is defined as:

$$G = (\mathbf{x}, \mathbf{v}, \varphi, w, d), \quad G \in \mathbb{R}^9 \tag{1}$$

where $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{R}^3$ is the grasp center position, $\mathbf{v} = (v_1, v_2, v_3) \in \mathbb{R}^3$ is the grasp approach vector, $\varphi \in \mathbb{R}$ is the rotation angle of the grasper's plane, $w \in \mathbb{R}^+$ is the grasp width, and $d \in \mathbb{R}^+$ is the grasp depth.

We define the grasp quality function $Q(P, G)$, where a higher value indicates a more stable grasp. In addition, to incorporate collision avoidance as a constraint, we define the collision function $C(P, G)$ as follows:

$$C(P, G) = \sum_{i=1}^{n} \mathbf{1}_{f_\theta(p_i, G) > \epsilon} \tag{2}$$

where $f_\theta(p_i, G)$ is a function that predicts whether point p_i is in collision with the grasper at pose G, ϵ is a predefined

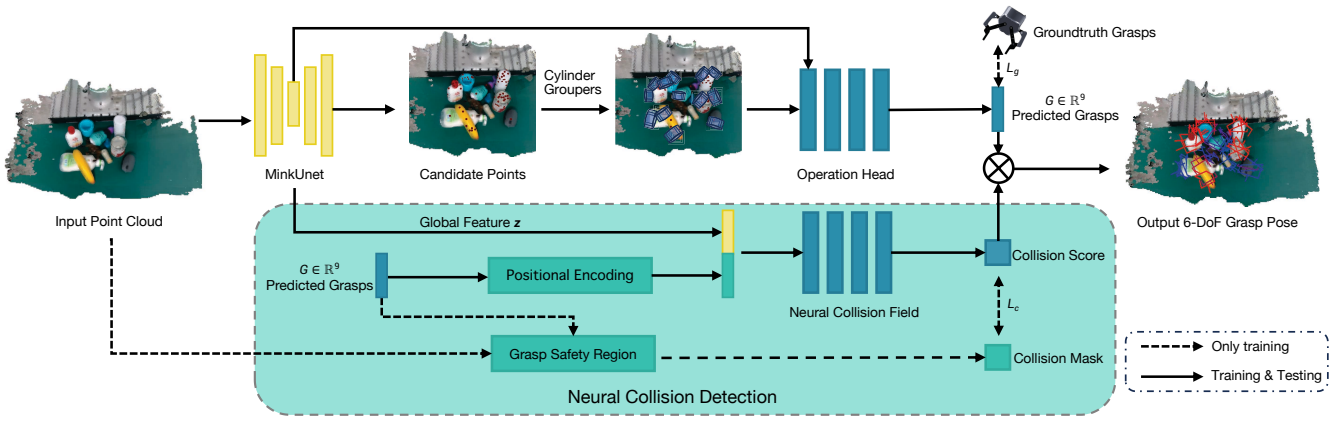


Fig. 1. Overview of our collision-aware 6-DoF grasp detection framework, which consists of two main modules. The 6-DoF grasp detection network processes the fused point cloud using a MinkUNet backbone to extract global and point-wise features, enabling the sampling of candidate points. Based on these, the Cylinder Graspers extract local features and fuse them with global features to predict grasp poses via the Operation Head. The Neural Collision Detection module predicts collision scores based on the global features and grasp parameters. The final grasp result is refined by integrating the predicted grasp poses with the collision scores, ensuring safe and efficient grasp execution.

collision threshold, set to 0 by default, and $\mathbf{1}_{\cdot}$ is an indicator function that returns 1 if a collision occurs at point p_i and 0 otherwise.

To ensure both a stable grasp and collision avoidance, we formulate the collision-constrained grasp pose detection problem as follows:

$$G^* = \arg\max_G \left(Q(P, G)\right) \tag{3}$$

subject to the constraint:

$$C(P, G) = 0 \tag{4}$$

This optimization seeks the grasp pose G^* that maximizes the grasp quality function while ensuring that no collisions occur. In cases where the constraint cannot be fully satisfied, we introduce a relaxed formulation as follows:

$$G^* = \arg\min_G \left((1 - Q(P, G)) + \lambda C(P, G)\right) \tag{5}$$

where λ is a weighting factor that balances grasp stability and collision avoidance. This formulation streamlines the problem, making it more amenable to solving through learning-based methods.

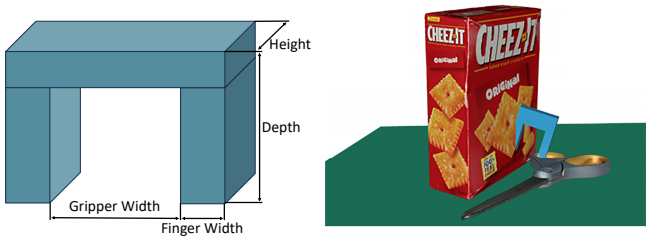
To address this problem, we employ neural networks to map point clouds P to the grasp pose G. As illustrated in the upper part of Fig. 1, we build on the work of [19] by leveraging ResUNet18 with MinkowskiEngine [28] as the backbone to efficiently process 3D point cloud data. The network first generates global features from the input point cloud, which are then used to predict the graspness score $s_{\text{graspness}}$, as described in [19]. These features help identify grasp candidate points $\mathbf{x}$ and their corresponding approach vectors $\mathbf{v}$. To refine these candidates, cylinder groupers sample local point clouds, which are processed by a local feature network to capture fine-grained details. These local features are subsequently fused with the global features, creating a multi-scale grasp representation. Finally, the fused features are passed to a module called the Operation Head, originally proposed in [20], which predicts the grasp parameters, including plane rotation φ, grasp depth d, grasp width w, and the final grasp quality score s_{grasp}. In the next section, we will introduce the implementation of the collision function $f_\theta(P, G)$, as shown in the lower part of Fig. 1.

B. Neural Collision Detection

To integrate collision detection into grasp pose optimization, we leverage implicit neural representations, which offer a continuous and differentiable framework for modeling collision functions. The proposed network architecture combines positional encoding $\gamma(\cdot)$ [29]–[31] with multilayer perceptrons (MLPs). Positional encoding maps 3D coordinates to a higher-dimensional space using high-frequency sinusoidal functions, which is effective for capturing fine-grained spatial details crucial for tasks like precise collision detection.

The collision detection network takes as input key grasp parameters from the grasp pose $G = (\mathbf{x}, \mathbf{v}, \varphi, w, d)$. The position vector $\mathbf{x}$, direction vector $\mathbf{v}$, and plane rotation angle φ are projected into the higher-dimensional space using positional encoding, and the grasp width w and depth d are directly included as input features. Global scene features $\mathbf{z}$, extracted from the 3D point cloud via ResUNet18, are also fed into the network. The neural collision function $f_\theta(\mathbf{z}, G)$ then estimates the likelihood of a collision-free grasp based on these inputs. Since the latent features $\mathbf{z}$ are encoded from the point cloud P, the $f_\theta(\mathbf{z}, G)$ approximates the collision function $f_\theta(P, G)$.

Training the network requires labeled data generated by simulating various grasping scenarios in 3D environments. To label this data, we define a spatial region called the Grasp Safety Region (GSR), which represents the potential collision space around the gripper. As shown in Fig. 2, the GSR is determined by the following parameters: finger width (w_f), height (h), and detection depth (d_{det}), where the detection depth is defined as $d_{det} = h + d$, with d representing the grasp



(a) (b)

Fig. 2. (a) illustrates the Grasp Safety Region (GSR) centered around the gripper. (b) presents an example of a collision occurring with a grasp.

depth. Specifically, the values are: $w_f = 0.01\,\text{m}$, $h = 0.02\,\text{m}$ as used in our experiments. The position and orientation of the GSR are directly determined by the grasp parameters G, and d is derived from the predicted grasp pose.

Each point in the scene is assessed to determine whether it falls within the Grasp Success Region (GSR). If it does, the corresponding grasp pose is labeled as being in collision (0); otherwise, it is considered collision-free (1). This classification defines the exact collision function $C(P, G)$, which serves as the foundation for a self-supervised labeling process.

The neural collision function is trained to minimize binary cross-entropy loss:

$$L_{collison} = -\frac{1}{n}\sum_{i}^{n} y_i log(f_\theta(\mathbf{z}, G_i)) + (1 - y_i) log(1 - f_\theta(\mathbf{z}, G_i)) \tag{6}$$

where y_i is the ground truth label derived from the GSR, i.e., $C(P, G)$. The Adam optimizer is used to minimize this loss during training. The output of the neural collision function is then integrated into the grasp pose optimization process.

During inference, the predicted collision-free probability $f_\theta(\mathbf{z}, G)$ is combined with the grasp scores from the previous grasp detection networks to refine the final grasp decision. The overall grasp score is calculated as follows:

$$s_{\text{total}} = s_{\text{grasp}} \times (1 - C(P, G)) \tag{7}$$

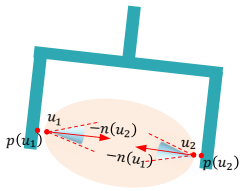
Here, s_{grasp} comes from the grasp detection network, representing the likelihood of a successful grasp. $C(P, G)$ represents the collision function, indicating the probability of collision for a specific grasp pose G within the scene. The final grasp score is then used to filter out undesirable grasp poses, ensuring that the selected grasp is both stable and free of collisions.

C. Regularizations

We now enhance our grasp optimization framework by introducing regularization terms that promote stability and feasibility in the predicted grasps. These regularizations are inspired by well-established physical and geometric principles in robotic grasping, namely force closure and form closure, which are essential for ensuring a stable and robust grasp.

1) Frictional Force Closure Condition: Grasp stability is determined by the interaction forces at the contact points between the gripper and the object, requiring these forces to satisfy force closure conditions. Traditional force closure frameworks [32] rely on precise object geometry but are computationally expensive, assume ideal conditions, and require accurate contact information, often unavailable in real-world scenarios.

Recent research [33] introduces the squeeze-force-closure (SFC) condition, offering a more practical approach to grasp stability. Unlike traditional methods that rely on detailed object geometry, SFC identifies antipodal contact points where forces are aligned oppositely along the surface normals, as illustrated in the figure. By considering the friction cones at these contact points, SFC effectively resists slippage and external disturbances, ensuring stable grasps even for irregularly shaped objects. This condition can be mathematically expressed as:



$$n(u_1) \cdot \frac{p(u_1) - p(u_2)}{|p(u_1) - p(u_2)|} > c_f \tag{8}$$

$$n(u_2) \cdot \frac{p(u_2) - p(u_1)}{|p(u_2) - p(u_1)|} > c_f \tag{9}$$

Here, μ represents the friction coefficient, $p(\cdot)$ denotes the contact point positions, and $n(\cdot)$ is the outward normal vector at each contact point. Unlike conventional approaches where $c_f = 1$, we introduce a physically grounded adaptation that adjusts c_f based on μ, allowing controlled deviations in the alignment between contact normals and the grasp axis:

$$c_f = \cos\left(\tan^{-1}\mu\right) \tag{10}$$

By linking c_f directly to the friction coefficient, our approach ensures grasp stability even when contact points experience different frictional forces. This makes the method more robust to real-world scenarios, where friction can vary due to surface texture and material properties.

To integrate the proposed frictional force closure (FFC) condition into the learning process, we define a loss function as follows:

$$L_{ffc} = \text{ReLU}\left(c_f - \cos\left(n(u_1), p(u_1) - p(u_2)\right)\right) + \text{ReLU}\left(c_f - \cos\left(n(u_2), p(u_2) - p(u_1)\right)\right) \tag{11}$$

where $cos(\cdot)$ represents the cosine similarity.

2) Symmetry Constraint: In addition to force-based stability, we impose a symmetry constraint to further stabilize the grasp by ensuring that the gripper's contact points are balanced relative to the object's centroid. This balance helps to distribute forces evenly across the grasp and reduces the likelihood of instability. The symmetry constraint is formulated as follows:

$$L_{sym}^k = \frac{1}{N}\sum_{i=1}^{N}\left|\left\|p_{\text{left}}^i - C^k\right\|_2 - \left\|p_{\text{right}}^i - C^k\right\|_2\right| \tag{12}$$

where C^k denotes the centroid of the object, p^i_{left} and p^i_{right} represent the left and right contact points of the i-th grasp pose, respectively. These points are calculated as follows:

$$p^i_{\text{left/right}} = R^i \left(d^i_{\text{det}},\ \pm\frac{w^i}{2},\ 0 \right)^\top + \mathbf{x}^i \tag{13}$$

where R^i transforms vectors from the local gripper frame to the global coordinate system. It is derived from the predicted approach vector $\mathbf{v}^i$ and the in-plane rotation angle φ^i. The parameters d^i_{det}, w^i, and $\mathbf{x}^i$ are obtained from the prediction of the i-th grasp pose $\mathbf{G}$.

This constraint encourages symmetry in the grasp configuration, promoting balanced force distribution and enhancing grasp stability.

3) Surface Alignment Constraint: We also recognize that predicted contact points may not always align perfectly with the object's surface, which could lead to unrealistic grasp predictions. To address this issue, we introduce a surface alignment constraint using the Signed Distance Function (SDF), which ensures that the contact points remain on the object's surface. This constraint is given by:

$$\begin{aligned} L_{sa} = \delta_1 \left(\text{ReLU}\left(d_1 - SDF(u_1)\right) + \text{ReLU}\left(d_1 - SDF(u_2)\right)\right) \\ + \delta_2 \left(\text{ReLU}\left(SDF(u_1) - d_2\right) + \text{ReLU}\left(SDF(u_2) - d_2\right)\right) \end{aligned} \tag{14}$$

where d_1 and d_2 are acceptable distance thresholds between the contact points and the object surface, and δ_1 and δ_2 are weighting factors. This constraint prevents excessive penetration or detachment, ensuring that the gripper interacts reliably with the object.

We integrate all the aforementioned principles into a unified framework, thus the regularizations are defined as:

$$L_{reg} = \alpha \left(L_{ffc} + L_{sym} \right) + \beta L_{sa} \tag{15}$$

where α and β are hyperparameters that control the influence of each term. This regularization ensures that the predicted grasps not only avoid collisions but also adhere to physically grounded stability criteria.

Finally, the total loss function used for training is defined as:

$$L = \lambda_0 L_{grasp} + \lambda L_{collision} + L_{reg} \tag{16}$$

where L_{grasp} is the grasp loss defined in [19], which evaluates the quality of predicted grasps, including objectness, graspability, and gripper width, using ground truth labels, i.e., $Q(P, G)$. The parameters λ_0 and λ are set to 1 by default. By incorporating collision-constrained grasp generalization, our approach guarantees that predicted grasps satisfy both physical and collision-free constraints, thereby resulting in stable and robust grasp configurations. This comprehensive strategy enhances the model's generalization capability, rendering it applicable across a diverse range of object geometries and manipulation scenarios.

IV. Experiments

A. Dataset and Model Training

1) Dataset and Benchmark: We evaluate our method on the widely-used GraspNet-1Billion dataset [18], which consists of 190 complex scenes: 100 for training and 90 reserved for testing. Each scene contains approximately 10 objects, totaling 3 to 9 million grasp candidates per scene. These grasps are captured using RealSense or Kinect cameras from various perspectives, resulting in 256 RGB-D images for each scene and corresponding 6D pose information. The test set is divided into three groups based on object categories: seen, similar, and novel.

For evaluation, we follow the metrics from the GraspNet-1Billion benchmark [18]. We modify the original grasp selection strategy, which ranks the top 50 grasp poses by score, to select the top k grasps, where $k = N_{\text{object}} \times 5$, with N_{object} being the number of objects in each scene. This adjustment allows for better handling of the complexity and diversity of grasping tasks in real-world scenarios.

2) Model Parameters: Our architecture is based on the 6-DoF grasp detection network presented in [19], [20], [34]. Specifically, we use ResUnet18 as the backbone for extracting point cloud features, leveraging the MinkowskiEngine [28] for efficient 3D convolutions. We adopt the multi-scale cylinder grouping technique from [20] to improve the extraction of local features.

For Neural Collision Detection network, we use 8 frequency components for position encoding, and the decoder MLP is structured with layers of sizes $(360, 512, 512, 256)$. In the formulation of the FFC condition (Equation 11), we set $c_f = 0.953$. Other parameters in Equation 14 and 15 are set as follows: $d_1 = 0.02$, $d_2 = 0.01$, $\delta_1 = \delta_2 = 1$, $\alpha = 0.1$, and $\beta = 0.05$.

3) Training Details: Our model is implemented in PyTorch, and we use the Adam optimizer for training. The experiments were conducted on a local server equipped with an Nvidia 4090 GPU (24 GB of memory), and the training process takes 320 iterations, approximating 7 hours to converge.

B. Performance Evaluation

We perform a comprehensive performance evaluation by comparing our model with several state-of-the-art methods [18]–[20], [26]. To assess grasping accuracy, we use the AP_μ metric proposed in [18], where AP_μ represents the average Precision@k for a given friction coefficient μ. The value of μ ranges from 0.2 to 1.2, with intervals of 0.2. This metric is averaged across these values of μ to provide a comprehensive evaluation of grasping performance.

We evaluated our model across three object categories: seen, similar, and novel. The results, summarized in Table I, demonstrate that our model consistently outperforms [26] across all categories. Specifically, our model achieved 67.75% AP on seen objects (a 3.68% improvement), 68.23% on similar objects (a 4.54% improvement), and 38.10% on novel objects (a 2.89% improvement). In terms of time efficiency, these accuracy improvements are achieved without compromising real-time performance. On the same hardware setup (identical GPU and CPU), our method processes point clouds at 17.5 FPS, which is comparable to [26] (17.9 FPS) and sufficient for real-world robotic deployment.

TABLE I

COMPARISON WITH THE STATE-OF-THE-ART METHODS ON REALSENSE SCENES OF GRASPNET-BILLION BENCHMARK.

Model	Seen			Similar			Novel		
	AP	$AP_{0.8}$	$AP_{0.4}$	AP	$AP_{0.8}$	$AP_{0.4}$	AP	$AP_{0.8}$	$AP_{0.4}$
GraspNet-baseline [18]	28.10	30.53	26.56	23.87	26.92	22.51	7.48	8.43	4.93
Scale-balanced Grasp [20]	46.05	51.02	44.27	37.76	44.27	34.75	17.09	21.04	10.20
GSNet [19]	60.47	71.05	52.06	58.55	69.38	53.08	28.06	36.19	14.09
Ma et al. [26]	64.07	72.54	60.61	63.69	72.84	59.91	35.21	41.75	21.28
Ours	**67.75**	**78.16**	**60.97**	**68.23**	**79.22**	**63.21**	**38.10**	**47.67**	**21.81**

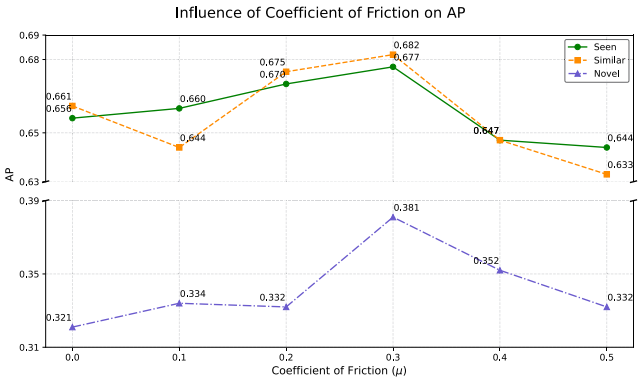


Fig. 3. The influence of c_f under different friction coefficients on the grasping results.

These results demonstrate the effectiveness of our approach in addressing diverse grasping scenarios while maintaining a practical balance between precision and processing speed.

C. The Influence of c_f on Grasping Performance

To further investigate the impact of friction on grasp performance, we analyze how variations in the friction coefficient μ influence the behavior of the FFC regularization. Existing work [26] assumes a constant value of $c_f = 1$, which enforces perfect alignment between the grasp axis and contact normals. However, real-world scenarios involve surface variations, which are better captured by dynamically adjusting c_f based on the friction coefficient μ.

We varied μ within the range of 0.0 to 0.5, with increments of 0.1, and analyzed the corresponding changes in c_f. The results, shown in Fig. 3, reveal that the AP for seen, similar, and novel objects peaks at $\mu = 0.3$. This suggests that a more relaxed c_f value allows for better adaptation to object surface variations, particularly improving the generalization of grasps, especially for novel objects.

D. Ablation Studies

To assess the individual contributions of each component of our method, we conducted ablation studies using the same 6-DoF Grasp Detection Network. The results of these experiments are presented in Table II, where we evaluate the effect of different regularization strategies on overall grasping performance.

Starting with the baseline model, which only uses the original grasp loss function from [19], we observe significant

TABLE II

ABLATION STUDIES OF OUR METHOD

Method	AP		
	Seen	Similar	Novel
Baseline (L_{grasp})	63.43	57.78	29.59
+ NCD ($L_{collision}$)	66.23	63.91	**35.77**
+ L_{reg}	**66.52**	**65.89**	35.66
+ L_{ffc}	66.12	64.80	34.65
+ L_{sym}	64.54	62.21	33.14
+ L_{sa}	65.23	63.91	32.77

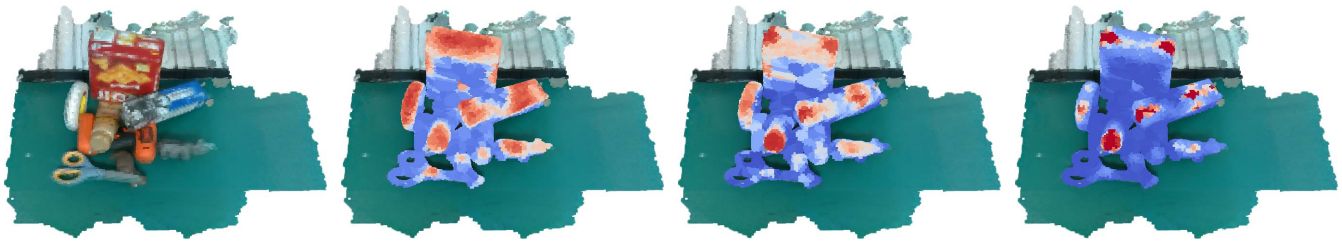
improvements with the addition of the Neural Collision Detection (NCD) module. Notably, the Novel AP increases from 29.59% to 35.77%, while the Seen and Similar APs also show improvement. The inclusion of the regularization terms further boosts performance, achieving the highest AP for both Seen (66.52%) and Similar (65.89%) objects, while maintaining a strong Novel AP of 35.66%.

Additionally, we evaluate the contribution of individual regularization modules. L_{ffc} achieves APs of 66.12%, 64.80%, and 34.65% for Seen, Similar, and Novel categories, respectively, while L_{sym} and L_{sa} provide incremental benefits, yielding 64.54% and 65.23% APs for Seen, 62.21% and 63.91% for Similar, and 33.14% and 32.77% for Novel, respectively. These results demonstrate that each module contributes to enhanced grasp detection, better generalization, and improved robustness.

E. Effects of Neural Collision Detection

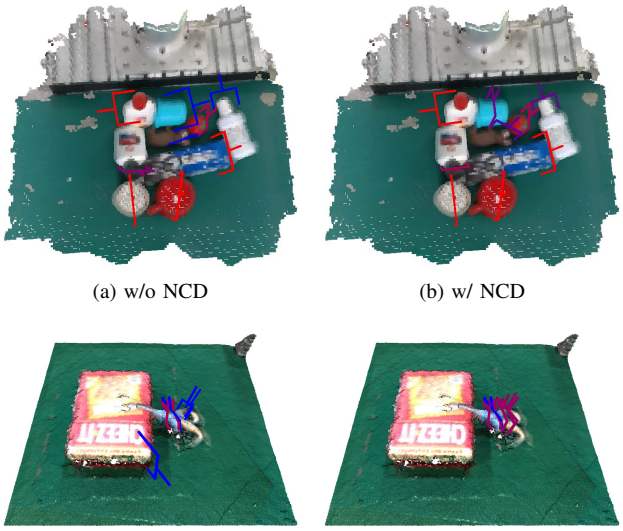
We further investigate the role of the Neural Collision Detection (NCD) in improving the feasibility of grasp poses. As shown in Fig. 4, the NCD module filters out high-risk grasps by prioritizing collision-free configurations. Figure 4b shows the distribution of Graspness scores [19], with higher likelihoods of successful grasps indicated in red. Figure 4c displays the predicted collision-free probability, highlighting poses with lower risk of collisions. Figure 4d integrates both Graspness score and collision-free probability leading to a more refined set of feasible grasp poses.

Further comparisons, as seen in Fig. 5, demonstrate the advantages of NCD: without NCD, numerous grasp candidates intersect with objects, increasing the likelihood of collisions (left column). In contrast, the inclusion of NCD (right column) refines the grasp proposals, focusing on stable and collision-aware poses. Quantitative results in Fig. 6 also show that NCD significantly reduces collision rates



(a) Input (b) Graspness score (c) Collision-free probability (d) b combined with c

Fig. 4. The visualization of the refined grasp pose selection process includes (a) the input scene, (b) the distribution of Graspness scores highlighting high-likelihood grasp regions in red, (c) NCD-predicted collision-free probabilities where higher red values indicate safer grasp poses, and (d) the combination of both scores, yielding a more focused and feasible distribution of viable grasp poses.



(a) w/o NCD (b) w/ NCD

(c) w/o NCD (d) w/ NCD

Fig. 5. Effect of NCD on grasp proposal refinement. Red indicates collision-free grasps, and blue indicates colliding grasps. (a) Without NCD, numerous grasp candidates intersect with objects, increasing the likelihood of collisions. (b) With NCD, collision-aware filtering enhances grasp feasibility by prioritizing collision-free poses. (c) Example without NCD, and (d) example with NCD, showing the impact of the filtering process.

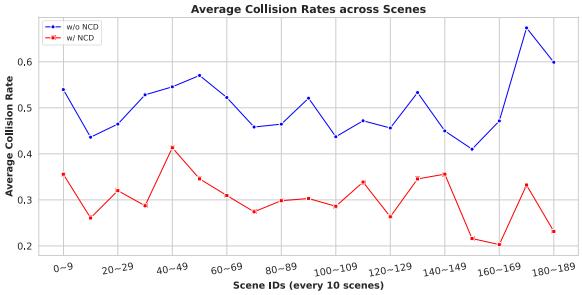


Fig. 6. Average collision rate across scenes with and without NCD. The proposed NCD (red) significantly reduces collision rates compared to the baseline without NCD (blue), demonstrating its effectiveness in improving collision-free grasping.

compared to the baseline, confirming its effectiveness in enhancing the collision-free performance of our method.

F. Real Robot Experiments

Finally, we validate our method in real-world grasping tasks using a UR10 robotic arm equipped with a Robotiq 2F-85 parallel gripper. The Intel RealSense D435i camera, mounted on the robot's wrist, captures depth images from multiple perspectives. These images are integrated into a Truncated Signed Distance Function (TSDF) via KinectFusion [35]. We conducted experiments with 30 items (Fig. 7(b)), including objects from the YCB dataset [36], 3D-printed items from Dex-Net [13], and daily-use objects.

(a) Robotic grasping system. (b) Objects for real experiments.

Fig. 7. (a) Robotic grasping system. (b) Objects used for real-world grasp experiments.

For each of the 20 scenes, containing 5 randomly placed items on a table, we performed 5 grasp attempts per scene. The success rate was calculated as the ratio of successfully grasped objects to the total number of attempts, resulting in an average success rate of 94% across all trials.

Furthermore, Fig. 8 illustrates the benefits of incorporating NCD in real-world experiments. Similar to the simulated environments in Fig. 5, without NCD, several grasp candidates overlap with objects; however, with NCD included, the grasp proposals are significantly refined.

V. Conclusion

This work introduced a collision-constrained grasp pose optimization framework that integrates a Neural Collision Detection network, effectively reducing grasp failures. By incorporating stability-enhancing regularization terms, our approach improves grasp feasibility, stability, and generalization to unseen objects. Experimental results demonstrate superior performance over state-of-the-art methods, highlighting the effectiveness of our method in grasp planning. These findings contribute to the development of more reliable robotic manipulation in cluttered environments.

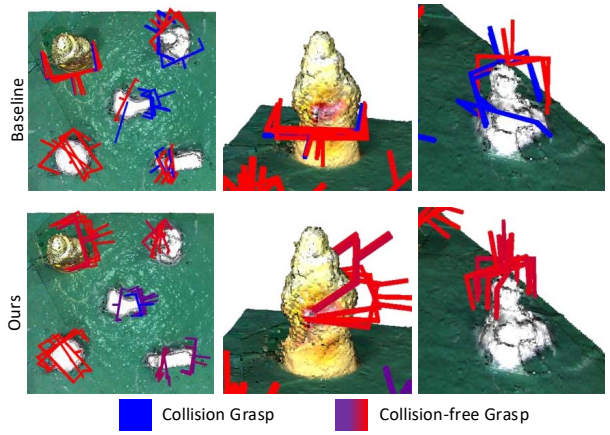


Fig. 8. Effect of NCD in real experiments, where red indicates a higher grasp quality score.

REFERENCES

[1] T. Pardi, R. Stolkin *et al.*, "Choosing grasps to enable collision-free post-grasp manipulations," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, 2018, pp. 299–305.

[2] X. Lou, Y. Yang, and C. Choi, "Collision-aware target-driven object grasping in constrained environments," in *2021 IEEE International Conference on Robotics and Automation*. IEEE Press, 2021, p. 6364–6370.

[3] L. Zhang, K. Bai, Q. Li, Z. Chen, and J. Zhang, "A collision-aware cable grasping method in cluttered environment," in *2024 IEEE International Conference on Robotics and Automation*, 2024, pp. 2126–2132.

[4] C. C. B. Viturino and A. G. S. Conceicao, "Selective 6d grasping with a collision avoidance system based on point clouds and rgb+ d images," *Robotica*, vol. 41, no. 12, pp. 3772–3787, 2023.

[5] J. Cai, J. Su, Z. Zhou, H. Cheng, Q. Chen, and M. Y. Wang, "Volumetric-based contact point detection for 7-dof grasping," *arXiv preprint arXiv:2209.06675*, 2022.

[6] V. Sitzmann, J. N. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Proc. NeurIPS*, 2020.

[7] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in neural information processing systems*, vol. 33, pp. 7537–7547, 2020.

[8] Y. Xie, T. Takikawa, S. Saito, O. Litany, S. Yan, N. Khan, F. Tombari, J. Tompkin, V. Sitzmann, and S. Sridhar, "Neural fields in visual computing and beyond," *Computer Graphics Forum*, 2022.

[9] S. Caldera, A. Rassau, and D. Chai, "Review of deep learning methods in robotic grasp detection," *Multimodal Technologies and Interaction*, vol. 2, no. 3, p. 57, 2018.

[10] G. Du, K. Wang, S. Lian, and K. Zhao, "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1677–1734, 2021.

[11] Z. Xie, X. Liang, and C. Roberto, "Learning-based robotic grasping: A review," *Frontiers in Robotics and AI*, vol. 10, p. 1038658, 2023.

[12] M. Dong and J. Zhang, "A review of robotic grasp detection technology," *Robotica*, vol. 41, no. 12, pp. 3846–3885, 2023.

[13] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," 2017.

[14] A. Ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.

[15] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang, "Pointnetgpd: Detecting grasp configurations from point sets," in *2019 International Conference on Robotics and Automation*, 2019, pp. 3629–3635.

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

[17] M. Breyer, J. J. Chung, L. Ott, R. Siegwart, and J. Nieto, "Volumetric grasping network: Real-time 6 dof grasp detection in clutter," in *Conference on Robot Learning*. PMLR, 2021, pp. 1602–1611.

[18] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[19] C. Wang, H.-S. Fang, M. Gou, H. Fang, J. Gao, and C. Lu, "Graspness discovery in clutters for fast and accurate grasp detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021.

[20] H. Ma and D. Huang, "Towards scale balanced 6-dof grasp detection in cluttered scenes," in *Conference on robot learning*. PMLR, 2023, pp. 2004–2013.

[21] M. Niu, Z. Lu, L. Chen, J. Yang, and C. Yang, "Vergnet: Visual enhancement guided robotic grasp detection under low-light condition," *IEEE Robotics and Automation Letters*, vol. 8, no. 12, pp. 8541–8548, 2023.

[22] H. Nie, Z. Zhao, L. Chen, Z. Lu, Z. Li, and J. Yang, "Smaller and faster robotic grasp detection model via knowledge distillation and unequal feature encoding," *IEEE Robotics and Automation Letters*, 2024.

[23] Z. Zhou, X. Zhu, and Q. Cao, "Aagdn: Attention-augmented grasp detection network based on coordinate attention and effective feature fusion method," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3462–3469, 2023.

[24] A. D. Vuong, M. N. Vu, H. Le, B. Huang, H. T. T. Binh, T. Vo, A. Kugi, and A. Nguyen, "Grasp-anything: Large-scale grasp dataset from foundation models," in *2024 IEEE International Conference on Robotics and Automation*, 2024, pp. 14 030–14 037.

[25] N. Nguyen, M. N. Vu, B. Huang, A. Vuong, N. Le, T. Vo, and A. Nguyen, "Lightweight language-driven grasp detection using conditional consistency model," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2024, pp. 13 719–13 725.

[26] H. Ma, M. Shi, B. Gao, and D. Huang, "Generalizing 6-dof grasp detection via domain prior knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 102–18 111.

[27] Z. Chen, Z. Liu, S. Xie, and W.-S. Zheng, "Grasp region exploration for 7-dof robotic grasping in cluttered scenes," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023, pp. 3169–3175.

[28] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[29] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[30] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[31] Y. Zhuang, "A simple and effective filtering scheme for improving neural fields," *Computational Visual Media*, vol. 11, no. 2, pp. 343–359, 2025.

[32] V.-D. Nguyen, "Constructing force-closure grasps," *The International Journal of Robotics Research*, vol. 7, no. 3, pp. 3–16, 1988.

[33] I.-M. Chen and J. W. Burdick, "Finding antipodal point grasps on irregularly shaped objects," *IEEE transactions on Robotics and Automation*, vol. 9, no. 4, pp. 507–512, 1993.

[34] D. Wei, J. Cao, and Y. Gu, "Robot grasp in cluttered scene using a multi-stage deep learning model," *IEEE Robotics and Automation Letters*, 2024.

[35] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE international symposium on mixed and augmented reality*, 2011, pp. 127–136.

[36] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*, 2015, pp. 510–517.