2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

# 6-DoF Grasp Detection in Clutter with Enhanced Receptive Field and Graspable Balance Sampling

Hanwen Wang[1], Ying Zhang*[1], Yunlong Wang[2] and Jian Li[1]

***Abstract*— 6-DoF grasp detection of small-scale grasps is crucial for robots to perform specific tasks. This paper focuses on enhancing the recognition capability of small-scale grasping, aiming to improve the overall accuracy of grasping prediction results and the generalization ability of the network. We propose an enhanced receptive field method that includes a multi-radii cylinder grouping module and a passive attention module. This method enhances the receptive field area within the graspable space and strengthens the learning of graspable features. Additionally, we design a graspable balance sampling module based on a 3D segmentation network, which enables the network to focus on features of small objects, thereby improving the recognition capability of small-scale grasping. Our network achieves state-of-the-art performance on the GraspNet-1Billion dataset, with an overall improvement of approximately 10% in average precision@k (AP). Furthermore, we deployed our grasp detection model on pybullet grasping platform and in real-world scenarios, which validates the effectiveness of our method.**

## I. INTRODUCTION

In recent years, grasp tasks for robotic arms have attracted significant attention in the fields of computer vision and deep learning [1]. In the execution of grasping tasks by robots, grasp detection serves as a fundamental task, providing the robot with perceptual capabilities for the scene. The goal of grasp detection is, given a scene containing objects, to identify a set of grasp configurations (including grasp point locations, joint poses, etc.) where the robotic hand, when closed at that configuration, can robustly grasp the corresponding object. Traditional grasp detection methods are primarily model-based [2]. These methods generate multiple grasp poses satisfying stability conditions based on the 3D model of the object. Subsequently, they calculate grasp poses satisfying stability conditions after coordinate transformations based on the actual pose of objects in the scene, thereby completing the grasping process. However, this approach heavily relies on the accuracy of object pose estimation in the scene. With the advancement of deep learning technology and the reduction in the cost of depth-sensing devices like Kinect and RealSense, data-driven grasp detection methods leveraging deep learning have steadily gained popularity [3].

*Corresponding author.
[1]Hanwen Wang, Ying Zhang and Jian Li are with the School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications, Beijing, 100876, China. {whw2022111391, yingzhang_bupt}@bupt.edu.cn, jianli_628@126.com
[2]Yunlong Wang is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, china. yunlong.wang@cripac.ia.ac.cn

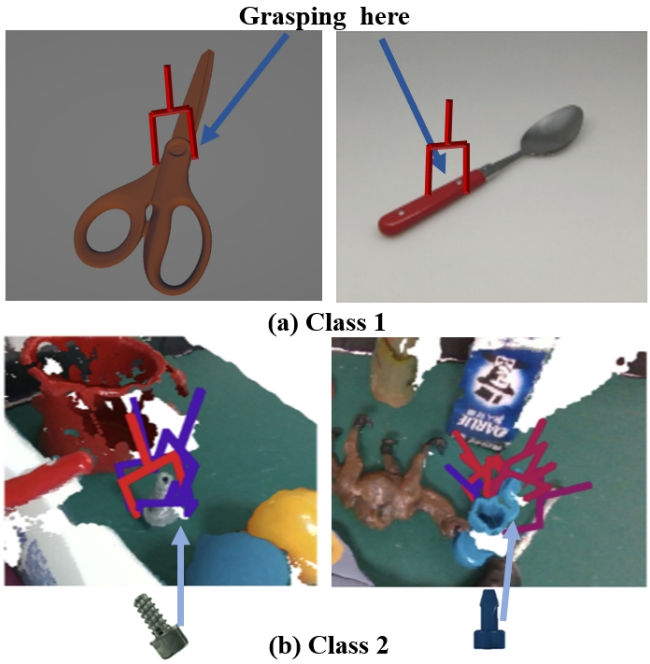



Fig. 1. We categorize small-scale grasping into two classes. The first class involves grasping small parts of medium to large objects on the tabletop. The second class pertains to grasping small objects at the tabletop level.

Deep learning-based grasping can be categorized into planar grasping and 6-Degree-of-Freedom (6-DoF) grasping [1]. Research on planar grasping detection primarily utilizes RGB or depth maps (RGB-D) as input and predicts a set of rotated bounding boxes to represent the grasping poses [4], [5], [6], [7], [8]. Due to the significant limitation of planar grasping, which confines the grasping poses to vertical motions from top to bottom, this method faces challenges when operating in complex real-world scenarios. 6-DoF grasping, designed for more versatile scenarios, offers greater flexibility compared to planar grasping. 6-DoF grasping detection networks leveraging deep learning technologies can predict the 6 degrees of freedom of the grasping pose, along with the opening width of the two-finger gripper. Earlier methods often employed a two-step sampling and evaluation approach, such as GPD [9] and PointnetGPD [10]. Due to the typically low quality of sampled results, a large number of samples need to be evaluated, leading to considerable time consumption. With the development of grasping detection datasets, end-to-end networks are easier to design and can fully exploit the information inherent in the data. Fang et al. [11] proposed the GraspNet-1Billion grasping detection

dataset, which has spurred various 6-DoF grasping detection research [12], [13], [14], [15]. However, previous methods still have deficiencies in detecting small-scale grasps, Fig.1 illustrates the concept of small-scale grasping. The uneven distribution of grasp pose samples across different scales in the dataset leads to the deteriorated recognition performance for small-scale grasps [15], thereby affecting the overall performance of the detection network. Although GSNet [14] annotates the feasible grasp space, directly sampling in this space may easily overlook features beneficial for regressing small-scale grasps. Similarly, subsequent parts using single-scale cylinder grouping may encounter similar issues. This issue hinders the smooth execution of task-oriented grasping operations by robots.

In this paper, we present a novel 6-DoF grasp detection network. Firstly, we propose enhanced receptive field method, which use Multi-radii Cylinder Grouping (MrCG) module to increase the receptive field area in the graspable space, and use Passive Attention (PA) module to enhance the sampled features. This enhancement facilitates better perception of fine details in small and medium-to-large objects. The feature based on graspable points also contributes to improved guidance for grasp pose prediction. Secondly, we introduce an 3D segmentation network based graspable balance sampling module. This module utilizes a pre-trained point cloud segmentation network to obtain the category of each point. Subsequently, it performs balanced sampling of graspable points learned by the network on each object, ensuring an equal number of sampled points for each object, therefore, small objects receive adequate attention from the model. Leveraging the 3D segmentation network, this method provides a solution for more advanced grasp tasks. Our approach outperforms previous state-of-the-art methods by 10% in AP on the GraspNet-1Billion dataset. Furthermore, in terms of the evaluation benchmark based on grasp scale, our method surpasses other approaches. Finally, we construct a 6-DoF grasp platform using pybullet tools and real-world robotic arm to conduct grasp tests on our detection network. Experimental results demonstrate the precision of our method in accomplishing 6-DoF grasp tasks in clutter scenes. Our contribution can be summarized as follows:

(1) We propose a enhanced receptive field method, which increases the model's receptive field area, enhancing the perception of small-scale graspable features.
(2) We propose an 3D segmentation network based graspable balance sampling method. This method improves the grasp detection network's perception of fine details in small and medium-to-large objects. Leveraging an segmentation network, this method provides a solution for more advanced grasp tasks.
(3) We validate our 6-DoF grasp detection method on the pybullet platform and real-word system. The experiments demonstrate that our method can accurately perform 6-DoF grasp tasks in clutter scenes.

## II. RELATED WORK

Our focus is primarily on reviewing 6-DoF grasp detection methods based on deep learning. These methods can be broadly categorized into two types: the first category involves sampling and evaluation-based methods, while the second category comprises end-to-end network methods.

**Sampling and evaluation-based methods.** These methods initially generate multiple grasping poses for a given scene. Subsequently, evaluate the sample according to a quality estimation function, fulfilled by a deep neural network [16], [17], [18], [9], [10]. Some methods further employ optimization-based approaches to refine the samples and generate higher-quality grasp poses [19], [20], [21], [22]. A major drawback of sampling and evaluation methods is the need to strike a balance between computation time and the quantity of generated grasp poses. As a result, these methods typically require several seconds to run and can only generate dozens of grasp poses for a single scene. With the development of grasp detection datasets, the advantages of end-to-end networks have gradually become evident. The increased availability of data allows for leveraging the power of representation learning of end-to-end networks, which are easy to design, effectively utilize the information present in the data itself, and exhibit fast inference speeds.

**End-to-end network methods.** These methods can directly regress grasping poses on the scene. Early literatures use RGB-D inputs to generate grasp poses [12], [23], [24]. Fang et al. [11] introduced a large-scale 6-DoF grasp dataset and proposed an inference network based on point cloud deep learning to regress dense grasp poses on the scene. Wang et al. [14] presented an evaluation method based on "graspness" to learn graspable regions. Ma et al. [15] investigated the issue of grasp scale imbalance in the GraspNet-1Billion dataset, identifying problems with the annotated grasp scales and proposing solutions. Breyer et al. [25] introduced a Volume Grasping Network (VGN), which takes the Truncated Signed Distance Function (TSDF) representation of the scene as input and outputs predicted grasp quality, fixture direction, and opening width for each voxel in the queried 3D volume. Dai et al. [26] proposed a 6-DoF grasp detection network, GraspNeRF, based on multi-view RGB inputs, addressing the problem of 6-DoF grasp detection for transparent and reflective objects.

However, grasp detection still faces challenges in terms of generalization and precision. Additionally, the detection capabilities for fine details in medium-to-large objects and small objects remain limited, leading to poor grasp pose availability. These challenges impede task-oriented grasping research.

## III. METHOD

In this section, we first briefly introduce the pipeline overview of grasp detection in clutter with enhanced receptive field and graspable balance sampling, as shown in the Fig.2. Then, we will focus on introducing our enhanced receptive field method and discuss two constituent modules:

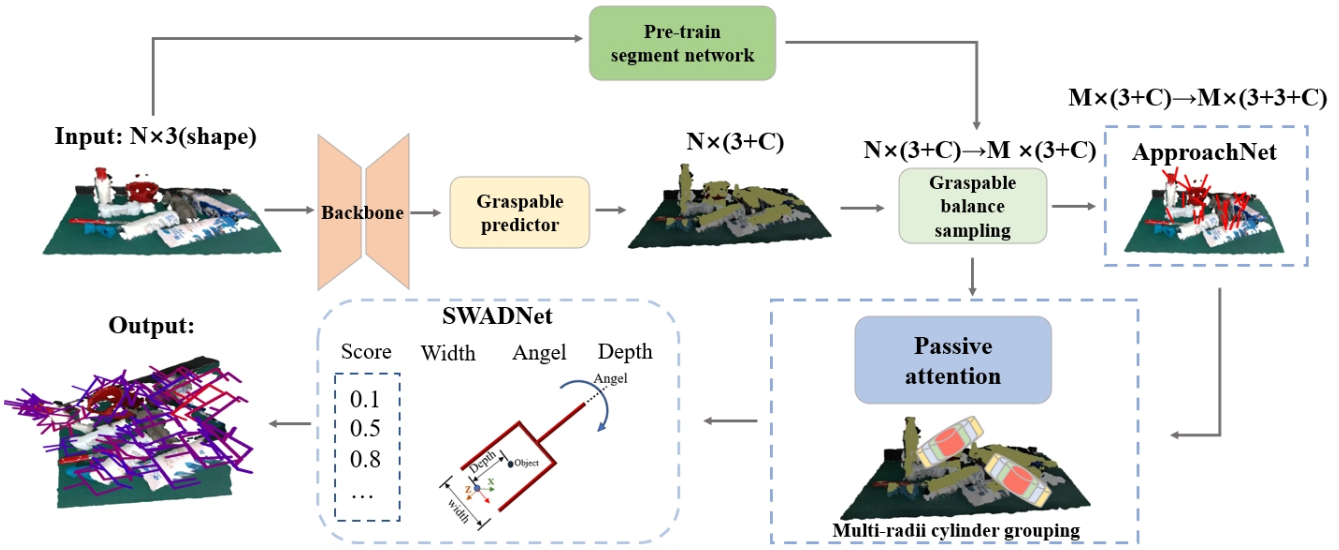



Fig. 2. Pipeline. The network initially generates multiple features through the backbone, followed by the graspable predictor predicting points with high graspness. The graspable balance sampling module has two modes: during training, it directly uses the farthest point sampling without employing the guidance of pre-trained segmentation model features. The model-guided sampling is utilized only during inference. The features are then fed into the ApproachNet to select the optimal grasp views, which is subsequently input into our enhanced receptive field for cylinder grouping. Finally, the input is processed by SWADNet to output dense grasp poses.

multi-radii cylinder grouping module and the passive attention module. Finally, we will present the graspable balance sampling module based on a pre-trained segmentation network, which exhibits two different forms during training and inference.

### A. Pipeline Overview

Our grasp detection network drew inspiration from [14], where they framed the grasping problem as a Bayesian problem involving the localization of grasp points and how to grasp, $X$, $Y$, $Z$ represent the grasp coordinates, $V$ represents the approach vector, $R$ represents the grasp rotation angle, $D$ represents the grasp depth, and W represents the width of the gripper opening. which is illustrated in Fig.3. $X$, $Y$, $Z$, and $V$ represent where to grasp, while $R$, $D$, and $W$ represent how to grasp, as illustrated in the formula:

$$P(Grasp) = P(R,D,W|X,Y,Z,V)P(X,Y,Z,V) \qquad (1)$$

We believe that learning based on graspable regions can enhance the accuracy of robotic grasp recognition. Initially, it is necessary to annotate spatial-level grasp ability in scenes of the dataset. Here, we briefly introduce the point-wise graspness score denoted as $\tilde{s}_i^{\mathrm{P}}$. This score involves sampling multiple grasp poses based on points in the scene, followed by an evaluation using force analysis conditions [27], [7] and scoring as $q_k^{i,j}$, and $\mathbf{1}(\cdot)$ is used to predict whether any grasp pose in space is valid. Grasps with collisions are filtered out. Indices $i$ and $j$ represent the point and approach vector, respectively. Index $k$ represents the grasp candidate $\mathscr{G}_{i,j}$, $L$ denotes the number of grasp candidates, $c$ is a scoring threshold, and $c_k^{i,j}$ represents the collision label for grasps, $V$ represents the numbers of view, the simplified version of

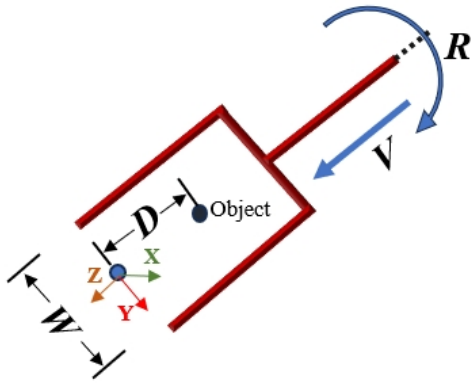



Fig. 3. Grasp representation and gripper coordinate system.

the formula is as follows:

$$\tilde{s}_i^p = \frac{\sum_{j=1}^{V}\sum_{k=1}^{L}\mathbf{1}(q_k^{i,j} > c)\cdot\mathbf{1}(c_k^{i,j})}{\sum_{j=1}^{V}|\mathscr{G}_{i,j}|}, \quad i = 1,\ldots,N \qquad (2)$$

Our innovation lies in proposing two methods to enhance the perceptual capabilities of small-scale grasping in the graspable space, as Fig.2 shows. Firstly, during the training phase of the network, we modify the cylinder grouping by using multiple cylinders to sample features for a single graspable point. We believe this method can yield a more powerful receptive field, reinforcing the accuracy of graspable point prediction in the earlier stages. Secondly, in the pure inference phase, we utilize a pre-trained point cloud segmentation network to guide the scene-level graspable point sampling, ensuring a balanced collection of graspable points on objects. Finally, our network can estimate denser grasp poses on the scene, where information about small-scale grasps may

(a) Single-radii cylinder grouping on graspable scene. (b) Multi-radii cylinder grouping on graspable scene.

Fig. 4. Schematic diagram of the cylinder grouping module. On the left is the conventional single-radii module, and on the right is our multi-radii module.

provide clues for future higher-level operations. Below, we will focus on elaborating the details of our methods.

### B. Enhanced Receptive Field

As shown in the Fig.4, previous works often used refined sampling points as the centers when performing cylinder grouping, setting a fixed radii for cylinder grouping [11], [14]. The consequence of this approach is a severe limitation on the receptive field area, leading to the failure of deep neural networks to learn the features of small parts of medium-to-large objects and grasp points of small-sized objects in clutter. To enhance the predictive capability for small-scale grasps and improve the network's regression performance, the network needs to consider more detailed geometric information. In this paper, we utilize the cylinder grouping method and enhance the receptive field after grasp sampling. as shown in the formula:

$$\mathscr{C}_q = \{\mathbf{v}_{ij} | \|p_{ij} - p_{ij_k}\| \leq r_q\}, \quad q = 1, \ldots, 4. \tag{3}$$

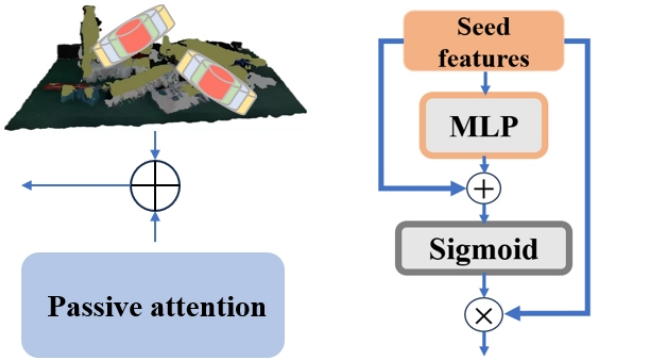
Where $\mathscr{C}_q$ represents the result of a single cylinder grouping. This method clusters by limiting the radii $r_q$. The key difference from previous work lies in our use of four different radii. Meanwhile, the radii of these four sampled cylinders uniformly increase within the maximum width of the gripper of the manipulator. Finally, the multi-radii clustering groups are processed with :

$$\mathscr{C} = Concat\{\mathscr{C}_1, \mathscr{C}_2, \mathscr{C}_3, \mathscr{C}_4\}. \tag{4}$$

Simultaneously, this module requires a passive attention module to guide the fused features, as shown in the Fig.5. This module is trained using features sampled from the grasp scene. The reason for incorporating this module may be the introduction of noise due to the increased receptive field. We believe that performing multi-scale grouping on the basis of graspable scenes can strengthen the predictive capabilities of the grasp network and improve its generalization performance. Although an increased receptive field may lead to noise perception, the robustness of the network against interference is enhanced in graspable scenes, as the graspness score in these scenes is continuous. This approach deepens the network's understanding of scene details, allowing for a comprehensive perception of geometric features in small parts.

### C. Graspable Balance Sampling

Previous research utilized a point cloud segmentation network to fully leverage the capabilities of farthest point sampling (FPS), ensuring an equal number of points are sampled on each object [15]. However, they did not conduct sampling under conditions that reasonably position the grasping space. This approach fails to guarantee that the sampled points do not contain excessive noise, which may hinder the overall grasp prediction performance, resulting in the prediction of low-quality grasps.




(a) Enhanced receptive field details (b) Passive attention module details

Fig. 5. On the left is our designed enhanced receptive field method. On the right is the pipeline of our passive attention module.

**Algorithm 1:** Balanced Sampling in the Graspable Space

**Input :** N points
**Output:** M points

$N_{ps} \leftarrow M/idx$
**for** $j \leftarrow 1$ **to** $idx$ **do**
  **if** $N_g = 0$ **then**
    Use FPS on the object points;
  **else if** $N_g < N_{ps}$ **then**
    Sample all graspable points on the object, then supplement with FPS other points on the object;
  **else**
    Use FPS on the object graspable points;
  **end**
**end**

To address this issue, we propose an 3D segmentation network based graspable sampling method, which use GBS module for inference. During the grasp pose inference phase, we utilize a pre-trained lightweight point cloud segmentation network [28]. Due to segmentation, we can effectively identify the class ownership of the point cloud in the scene, enabling us to commence balanced sampling. The details of our algorithm are described in Algorithm 1. Point clouds are acquired from both graspable scenes and original scenes. Our goal is to sample more graspable points as inputs for our network. However, on certain objects, it may be challenging to predict enough graspable points. Firstly, The number of grasp points needed for each object is calculated. Then, we set up three sampling scenarios. In the first scenario, when

TABLE I

ABLATION STUDY OF THE PROPOSED MODULES RESULTS ON REALSENSE D435

| Model | Seen | | | Similar | | | Novel | | |
|---|---|---|---|---|---|---|---|---|---|
| | **AP** | **AP**$_{0.8}$ | **AP**$_{0.4}$ | **AP** | **AP**$_{0.8}$ | **AP**$_{0.4}$ | **AP** | **AP**$_{0.8}$ | **AP**$_{0.4}$ |
| w/o PA | 73.46 | 85.13 | 68.33 | 64.21 | 77.02 | 56.38 | 26.63 | 33.19 | 14.04 |
| w/o MrCG | 69.92 | 81.51 | 64.27 | 60.99 | 73.93 | 52.22 | 25.70 | 32.13 | 13.67 |
| Ours | 74.33 | 85.77 | 63.89 | 64.36 | 76.76 | 55.25 | 27.56 | 34.09 | 20.23 |

there are no graspable points on the object, we perform FPS on the original point cloud in the scene. In the second scenario, if the number of graspable points on the object is less than the required sampling points, all graspable points on the object are sampled first, and then FPS is performed on the remaining points. In the third scenario, when the number of graspable points on the object meets the sampling point requirement, FPS is directly performed on the graspable points of the object. Simultaneously, our method can more efficiently regress semantic-based grasp poses in everyday usage scenarios. This approach provides rich visual clues for more advanced task operations.

## IV. EXPERIMENTS

### A. Implementation Details

**Benchmark dataset and metric.** Our network is trained and tested on the GraspNet-1Billion dataset [11], which includes 190 scenes. Each scene contains information captured from 256 viewpoints, and dense grasp poses are annotated for each scene. The test set is divided into three categories based on difficulty: seen, similar, and novel. We employ two evaluation methods to assess our network. Firstly, we use precision@k as our evaluation metric, which measures the precision of the top k-ranked grasps. **AP**$\mu$ represents the average precision@k under a given friction coefficient $\mu$. This evaluation method utilizes dynamic force closure analysis, which aligns better with real-world grasp success conditions. Secondly, we aim to validate that our network contributes to improving the recognition performance of small-scale grasps. We adopt the method proposed in [15], which defines the grasp scale as the gripper's opening width and categorizes it into three classes: widths in 0cm-4cm, 4cm-7cm, and 7cm-10cm as small-scale, medium-scale, and large-scale. Following the same dynamic force closure analysis, we evaluate the scene's $\mathbf{AP}_S$, $\mathbf{AP}_M$, and $\mathbf{AP}_L$.

**Network Implementations**. Our network implementations involve the use of 4D convolutions [29] in the backbone to extract and learn features. The output features are increased to 512 channels. The graspable predictor in our network employs MLP for predicting the graspable positions. The shape of the MLP for learning grasp points is (512, 3), and for learning viewpoints, it is (512, 3). In the training phase for grasp point sampling, FPS is directly used in the graspable space, while in the inference phase, our segmentation network based balance graspable sampling method is employed for sampling. In the enhanced receptive field stage, we perform multi-scale cylinder sampling using the sampled points and features, with four different radii (0.0125m, 0.025m, 0.0375m, 0.05m), then concatenate the features of the four groups and pass them through a MLP with a shape of (1024, 512). In the PA module, the shape of the MLP is (512, 512). SWADNet is used for yielding final grasp rotation angles and grasp depth, and the shapes of the two MLPs inside it are (512, 256) and (256, 48).

TABLE II

ABLATION STUDY OF THE PROPOSED MODULES IN SMALL SCALE GRASPING RESULTS ON REALSENSE D435

| Model | Seen | Similar | Novel |
|---|---|---|---|
| GSNet[14] | 20.07 | 6.75 | 9.59 |
| Ma et al. [15] | 18.29 | **10.03** | 9.29 |
| w/o PA | 22.11 | 8.82 | 11.36 |
| w/o MrCG | 21.22 | 7.64 | 10.76 |
| Ours | 23.01 | 8.89 | 11.33 |
| Ours + GBS | **23.67** | 9.21 | **11.38** |

**Training and Inference**. Our model is implemented with PyTorch and trained on one NVIDIA Tesla V100 GPU for 13 epochs with Adam optimizer [30] and the batch size of 4. The learning rate is 0.001 at the first epoch, and multiplied by 0.95 every one epoch. The network takes about 2 day to converge. In inference with collision detection, we only use one GPU for prediction.

### B. Ablation Study

In this section, we primarily validate the effectiveness of our proposed MrCG and PA modules on the GraspNet-1Billion dataset [11]. Initially, we employed the benchmarks [11] to evaluate these modules. As shown in the Table I, the MrCG module achieved a significant improvement in most evaluation metrics. Moreover, when MrCG and PA module were used in combination, there was a notable enhancement in the most stringent evaluation metric, $\mathbf{AP}_{0.4}$ by novel. This experiment demonstrates that our approach comprehensively improves grasp detection accuracy and exhibits certain generalization performance improvements compared to previous methods.

Secondly, our approach focuses on enhancing the recognition capability for small-scale grasps. We utilized [15] to propose an evaluation standard tailored to grasp scales to assess the model's ability to recognize small-scale grasps. We evaluated the effectiveness of the PA and MrCG modules in improving small-scale grasp recognition. Simultaneously, we assessed the effectiveness of using the GBS module for inference on small-scale grasps. The experiments indicate that there is a certain effect when using the GBS module for grasping. As shown in Table II, the MrCG module contributes the most, and compared to previous methods, we achieved state-of-the-art performance in the seen objects

TABLE III

ABLATION STUDY IN FULL SCALE GRASPING RESULTS ON REALSENSE D435

| Model | Seen | | | | Similar | | | | Novel | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $AP_S$ | $AP_M$ | $AP_L$ | Mean | $AP_S$ | $AP_M$ | $AP_L$ | Mean | $AP_S$ | $AP_M$ | $AP_L$ | Mean |
| GSNet[14] | 20.07 | 65.11 | 72.41 | 52.53 | 6.75 | 50.51 | 64.72 | 40.66 | 9.59 | 24.20 | 26.25 | 20.01 |
| Ma et al. [15] | 18.29 | 52.6 | 64.34 | 45.08 | 10.03 | 42.77 | 57.09 | 36.63 | 9.29 | 18.74 | 24.36 | 17.46 |
| Ours | 23.01 | 67.67 | 76.95 | 55.88 | 8.89 | 53.88 | 67.16 | 43.31 | 11.33 | 25.58 | 27.44 | 21.45 |
| Ours + GBS | 23.67 | 67.54 | 78.53 | **56.58** | 9.21 | 54.39 | 68.83 | **44.14** | 11.38 | 26.17 | 28.20 | **21.92** |

TABLE IV

GRASPNET-1BILLION EVALUATION RESULTS ON REALSENSE D435

| Model | Seen | | | Similar | | | Novel | | |
|---|---|---|---|---|---|---|---|---|---|
| | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ |
| GPD[9] | 22.87 | 28.53 | 12.84 | 21.33 | 27.83 | 9.64 | 8.24 | 8.89 | 2.67 |
| PointnetGPD[10] | 25.96 | 33.01 | 15.37 | 22.68 | 29.15 | 10.76 | 9.23 | 9.89 | 2.74 |
| GraspNet-baseline[11] | 27.56 | 33.43 | 16.95 | 26.11 | 34.18 | 14.23 | 10.55 | 11.25 | 3.98 |
| Gou et al. [12] | 27.98 | 33.47 | 17.75 | 27.23 | 36.34 | 15.60 | 12.25 | 12.45 | 5.62 |
| GSNet[14] | 67.12 | 78.46 | 60.90 | 54.81 | 66.72 | 46.17 | 24.31 | 30.52 | 14.23 |
| Ma et al. [15] | 63.83 | 74.25 | 58.66 | 58.46 | 70.05 | 51.32 | 24.63 | 31.05 | 12.85 |
| Ours | **74.33** | **85.77** | **63.89** | **64.36** | **76.76** | **55.25** | **27.56** | **34.09** | **20.23** |

and novel objects tests but slightly lower results in similar, possibly due to the use of cost-sensitive learning in [15], [31] approach. They designed a special loss function for weighting the errors of samples at different grasp scales, based on the frequency of grasp scale categories in the dataset.

Although our method focuses on improving the detection capability of small-scale grasps, our approach outperforms previous methods in grasp detection across all scales, as shown in Table III, this improvement is particularly significant when we employ GBS for grasp pose inference.

### C. Comparing with Representative Methods

We conducted comparisons with representative methods on the GraspNet-1Billion dataset. Methods utilizing point cloud input generally outperform those using RGB images, and our focus was on comparing with methods which take point cloud as input. We followed the testing methodology outlined in [11], conducting tests on three object categories. In our model, along with [14], and [15], we utilized collision detection for final predictions to achieve better results. The results of our comparative evaluation are reported in Table IV, We present only the RealSense section of the dataset because previous work has shown that training with Kinect data under the same network architecture results in suboptimal inference performance [11], whereas data collected using a RealSense camera yields better detection outcomes.

Our method achieved state-of-the-art performance compared to previous approaches. Using the FPS mothod for inference, our network showed an improvement of approximately 10% in AP metrics compared to previous methods. Notably, in the most challenging novel scenarios, our network's performance improved by 9.6% compared to previous methods.

### D. PyBullet Grasping Experiment

A platform for testing the grasping capability of small objects was established based on the pybullet robot tools [32]. Here are the details of our implementation. The Franka Panda gripper was utilized for grasping, aiming to validate the algorithm's effectiveness while conserving computational resources. To simulate the working space of a real manipulator, we filtered out grasping poses with an angle exceeding a threshold with the world coordinate system's Z-axis [33]. During the grasping process, the gripper was initially positioned to a pre-grasping pose, with the grasping pose set along the object's approach vector. When executing the grasp, the gripper moved along the approach vector. A successful grasp was defined as holding the object, moving to the pre-grasping position, and maintaining stability for a certain duration.

17 models are selected from the YCB dataset [34] and divided them into two categories, as shown in Table V, one for objects suitable for small-scale grasping poses, and the other for objects suitable for medium to large-scale grasping poses. We designed two types of experimental scenes based

TABLE V

TEST OBJECTS LIST IN YCB DATASET

| Objects categories | IDs |
|---|---|
| Small | 11, 12, 31, 37, 38, 72-i, 72-d, 72-f, 73-c |
| Medium to large | 4, 6, 10, 13, 16, 19, 21, 35 |

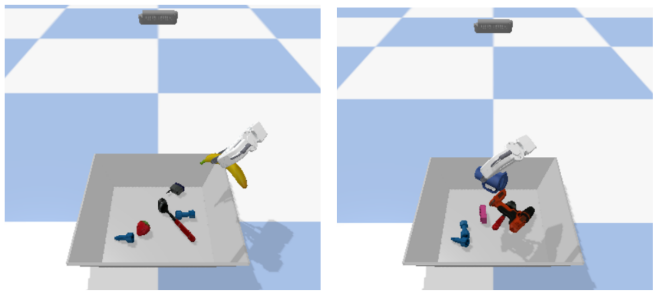


Fig. 6. We divided the scenes into two categories: the left category is small scale objects scene, and the right category is mixed scale objects scene.

TABLE VI

PYBULLET GRASPING RESULTS IN CLUTTERED SCENES

| Model | Small Scale objects Scenes SR | Mixed Scale objects Scenes SR |
|---|---|---|
| GSNet[14] | 0.86 | 0.84 |
| Ours | **0.95** | **0.87** |

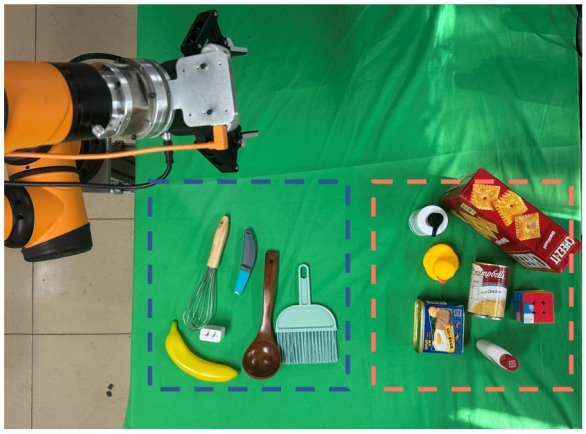


Fig. 7. Robot and object settings in real-world experiment: the blue box is used for testing small-scale grasps, while the red box contains objects for testing medium to large-scale grasps.

on object classification. As shown in Fig.6, the first type of scene contains only six small-scale objects. In the second type of scene, we randomly select six objects from the already chosen models to compose a new scene for grasping experiments. We refer to such scenes as mixed-scale objects scenes. It is worth mentioning that the objects in our scenes are randomly placed and cluttered, rather than being isolated.

As shown in Table VI, we conducted comparative experiments with the state-of-the-art methods in our pybullet environment, and tested our proposed model in two types of scenes, The scene completion rate of all our grasping tests is 1. Firstly, In the small-scale objects scenes, ours model attempted 63 grasps in total, achieving successful grasps 60 times, resulting in success rate (SR) of 0.95. GSNet attempted 70 grasps in total, achieving successful grasps 60 times, resulting in SR of 0.86. Our model gains an almost 10% improvement compared to GSNet [14]. Secondly, for mixed-scale objects scenes, we attempted 107 grasps in total, achieving successful grasps 90 times, resulting in SR of 0.87. GSNet [14] attempted 107 grasps in total, achieving successful grasps 90 times, resulting in success rate (SR) of 0.84. This experimental results validates that our method enhances the recognition capability of deep learning networks for small-scale grasping in clutter scenes, while also effectively addressing medium to large scale grasping recognition. Compared to previous methods, our model has achieved notable improvements in understanding objects grasping.

### *E. Real-world Grasping Experiment*

To verify the effectiveness of our grasp detection method in the real-world, we set up a physical scenario for grasping

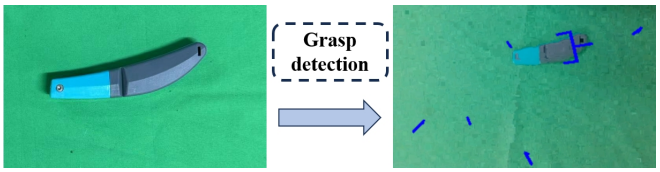



Fig. 8. Failure cases in real-word experiment

experiments. As shown in Fig.7, our grasping platform consists of an AUBO-i5 collaborative robot, a RealSense camera, a 2-Fingered Modular Changing Hand, and two categories of objects to be grasped. The grasping strategy and object deployment are similar to the simulation experiments, mainly testing our method's capability in small-scale and full-scale grasping. It is worth mentioning that in the experiments testing small-scale grasping ability, we selected some tools frequently used by humans in daily life. Evaluating our method on these objects is crucial because it needs to provide a foundation for generalized robotic manipulations in the future.

Our SR is 0.83 for small-scale grasping and 0.91 for mixed-scale grasping. This result is close to our success rate in simulation because the training datasets [11] was collected from real-world. However, it is challenging to capture the complete depth for flat objects. As shown in the Fig.8, when the grasp detection performance is less than optimal, our experimental method tends to choose some noisy poses. In such cases, using the highest grasp score is not suitable. One possible solution is to use image based instance segmentation methods to filter the grasp poses.

## V. CONCLUSIONS

In this paper, we propose an enhanced receptive field method, which increases the model's receptive field based on graspable space. This modification allows the model to pay more attention to subtle features on objects. Additionally, we propose a segmentation network based graspable balance sampling method. This method utilizes a point cloud segmentation model to extract points belonging to the grasped object, effectively filtering out noise. Graspable points on each object are uniformly sampled, ensuring that features of small objects are not overlooked and enhancing the recognition capability for small-scale grasping. The segmentation model also provides a solution for semantic grasping. We conducted extensive experiments to evaluate our proposed methods. The accuracy tests of the grasping detection network demonstrate that, compared to previous methods, our method improves the recognition ability for small-scale grasping at the visual level, enhancing overall network generalization and accuracy. Furthermore, objects grasping experiments confirm that our

grasping detection network can predict effective grasping poses. In the future, our work can provide grasp perception capabilities for task-oriented robotic manipulation.

## ACKNOWLEDGMENT

The research was partially supported by the National Natural Science Foundation of China (No. 52005120) and the Interdisciplinary Team of Intelligent Elderly Care and Rehabilitation in the "Double First-class" Construction of Beijing University of Posts and Telecommunications in 2023 (No. 2023SYLTD04).

## REFERENCES

[1] 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.

[2] H. Zhang, J. Tang, S. Sun, and X. Lan, "Robotic grasping from classical to modern: A survey," *arXiv preprint arXiv:2202.03631*, 2022.

[3] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, D. Fox, and A. Cosgun, "Deep learning approaches to grasp synthesis: A review," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3994–4015, 2023.

[4] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3355–3362, 2018.

[5] U. Asif, J. Tang, and S. Harrer, "Densely supervised grasp detector (dsgd)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8085–8093.

[6] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *2011 IEEE International conference on robotics and automation*. IEEE, 2011, pp. 3304–3311.

[7] 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," *arXiv preprint arXiv:1703.09312*, 2017.

[8] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *arXiv preprint arXiv:1804.05172*, 2018.

[9] 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.

[10] 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 (ICRA)*. IEEE, 2019, pp. 3629–3635.

[11] 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.

[12] M. Gou, H.-S. Fang, Z. Zhu, S. Xu, C. Wang, and C. Lu, "Rgb matters: Learning 7-dof grasp poses on monocular rgbd images," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 459–13 466.

[13] Y. Li, T. Kong, R. Chu, Y. Li, P. Wang, and L. Li, "Simultaneous semantic and collision learning for 6-dof grasp pose estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3571–3578.

[14] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 964–15 973.

[15] 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.

[16] J. Varley, J. Weisz, J. Weiss, and P. Allen, "Generating multi-fingered robotic grasps via deep learning," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 4415–4420.

[17] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 4304–4311.

[18] M. Gualtieri, A. Ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 598–605.

[19] Y. Zhou and K. Hauser, "6dof grasp planning by optimizing a deep learning scoring function," in *Robotics: Science and systems (RSS) workshop on revisiting contact-turning a problem into a solution*, vol. 2, 2017, p. 6.

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

[21] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2901–2910.

[22] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, "6-dof grasping for target-driven object manipulation in clutter," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6232–6238.

[23] J. Lundell, F. Verdoja, and V. Kyrki, "Ddgc: Generative deep dexterous grasping in clutter," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6899–6906, 2021.

[24] Y. Chen, Y. Lin, R. Xu, and P. A. Vela, "Keypoint-graspnet: Keypoint-based 6-dof grasp generation from the monocular rgb-d input," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 7988–7995.

[25] 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.

[26] Q. Dai, Y. Zhu, Y. Geng, C. Ruan, J. Zhang, and H. Wang, "Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1757–1763.

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

[28] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "Unseen object instance segmentation for robotic environments," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1343–1359, 2021.

[29] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3075–3084.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, no. 1. Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[32] D. Wang, F. Chang, C. Liu, H. Huan, N. Li, and R. Yang, "On-policy and pixel-level grasping across the gap between simulation and reality," *IEEE Transactions on Industrial Electronics*, 2023.

[33] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, 2023.

[34] 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)*. IEEE, 2015, pp. 510–517.