Sim-Suction: Learning a Suction Grasp Policy for Cluttered Environments Using a Synthetic Benchmark

Juncheng Li[1], David J. Cappelleri[1,2]

***Abstract*—This paper presents *Sim-Suction*, a robust object-aware suction grasp policy for mobile manipulation platforms with dynamic camera viewpoints, designed to pick up unknown objects from cluttered environments. Suction grasp policies typically employ data-driven approaches, necessitating large-scale, accurately-annotated suction grasp datasets. However, the generation of suction grasp datasets in cluttered environments remains underexplored, leaving uncertainties about the relationship between the object of interest and its surroundings. To address this, we propose a benchmark synthetic dataset, *Sim-Suction-Dataset*, comprising 500 cluttered environments with 3.2 million annotated suction grasp poses. The efficient *Sim-Suction-Dataset* generation process provides novel insights by combining analytical models with dynamic physical simulations to create fast and accurate suction grasp pose annotations. We introduce *Sim-Suction-Pointnet* to generate robust 6D suction grasp poses by learning point-wise affordances from the *Sim-Suction-Dataset*, leveraging the synergy of zero-shot text-to-segmentation. Real-world experiments for picking up all objects demonstrate that *Sim-Suction-Pointnet* achieves success rates of 96.76%, 94.23%, and 92.39% on cluttered level 1 objects (prismatic shape), cluttered level 2 objects (more complex geometry), and cluttered mixed objects, respectively. The codebase can be accessed at *https://github.com/junchengli1/Sim-Suction-API*.**

I. INTRODUCTION

THE development of autonomous mobile manipulation platforms is crucial for the future of space habitats, where robots can perform various tasks in cluttered environments with minimal human intervention. In these habitats, tasks such as maintenance, cargo handling, and assembly of structures have unique challenges for grasping and manipulation due to confined spaces, limited resources, and the need to handle objects with diverse shapes, sizes, and materials. Furthermore, space habitats often contain cluttered environments with objects that may be partially occluded or challenging to access. While space habitats represent a vital application area, mobile manipulation platforms also play an essential role in industry 4.0 and household settings due to their flexibility and efficiency. However, humans expect mobile manipulation platforms to be fully autonomous without any intervention. This is challenging for manipulation tasks, where robots have trivial or no pre-existing knowledge, unlike tasks on a predictable assembly line under controlled conditions.

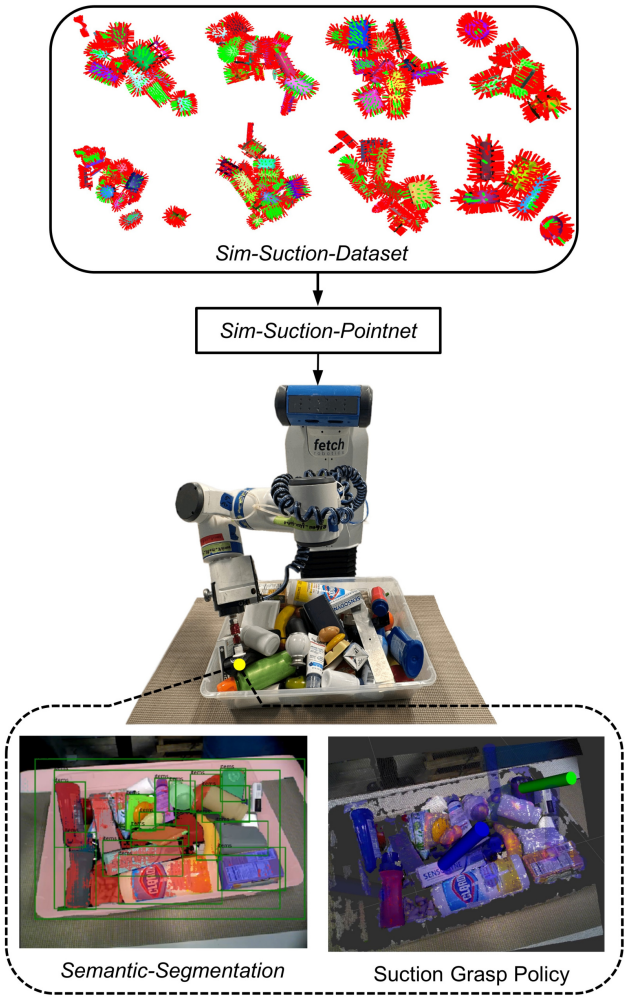


Fig. 1. Overview of *Sim-Suction*. The *Sim-Suction* is a deep-learning based policy to determine the robust suction grasp poses in cluttered environments. It has the following components: *Sim-Suction-Dataset*, a large-scale synthetic dataset for suction cup gripper that combines analytical model and physical simulation; *Sim-Suction-Pointnet*, an object-aware point-wise affordance network that uses text prompt to predict grasp success probability for given picking-up task.

For example, typical dynamic manipulation tasks may include picking objects from a bin, cleaning a cluttered workbench, or retrieving products from a shelf. These tasks are challenging for mobile manipulation platforms, both in terms of identifying the grasp region and executing the mechanical grasping process. Unlike familiar objects, novel objects are

[1] J. Li and D. Cappelleri are with the Multi-Scale Robotics & Automation Lab, School of Mechanical Engineering, Purdue University, West Lafayette, IN USA. [2] D. Cappelleri is also with the Weldon School of Biomedical Engineering (By Courtesy), Purdue University, West Lafayette, IN USA. {li3670, dcappell}@purdue.edu

items that the robot has never encountered before, hence, no prior information about their shape, size, weight, texture, and other physical properties is available. To solve these tasks, mobile manipulation platforms require the ability to observe the cluttered environments, decide on the way to grasp the object of interest, and perform a robust grasp once found. This is a difficult task due to the challenges associated with where to grasp and the uncertainties on how objects with varying size, weight, shape, and texture will react when trying to establish grasp contact point. The mobile manipulation platforms need to understand the task requirements that humans give and make the right decisions in such settings. Our work in this paper aims to tackle the picking-up challenges by using text prompts to guide the robot in completing tasks such as picking up all novel objects from a bin or selecting specific novel objects from the bin based on a brief text description.

Suction cup grippers play a vital role in warehouses due to their simplicity, compactness, light weight, and minimal maintenance requirements. They can also handle a wide range of objects, from fragile parts to large dimension objects. The Amazon Picking Challenge showed that the suction cup gripper is commonly used for general picking tasks with a higher success rate than other grippers [1]. Experiments from DexNet 4.0 [2] also demonstrate the preference for choosing a suction gripper over a parallel jaw gripper, with an 82% selecting rate on the bin-picking task. Previous studies show that the suction cup gripper outperforms other grippers in successfully grasping objects from cluttered environments due to its ability to create a single contact point on the object surface through a narrow space. Suction grippers are particularly suited for handling objects in space habitats with various surface properties, as they can establish a secure grip on a wide range of materials without causing damage. This is particularly important for delicate equipment and components that require gentle handling and precise placement. Our previous work designed a modular end-effector system [3], which enables a mobile manipulation platform to use a suction cup gripper more efficiently with an embedded vacuum generator and control module. In this work, we focus on developing the suction grasping policy for mobile manipulation platforms to tackle the challenge of grasping objects from a cluttered environment.

In the grasping research community, a large number of studies pay particular attention to the parallel-jaw gripper grasping policy. At the same time, a relatively small body of literature is concerned with the suction grasping policy. Studies on suction grasping [4]–[7] have begun to examine the data-driven approaches [8]–[11] using deep learning and achieve better performances over traditional online heuristic baseline approaches. The data-driven approaches can be classified into training on realistic datasets [5], [6] and training on synthetic datasets [4]. The key problem of realistic datasets collected from human or real experiments [12] is that the time cost to retrieve object information is large [13]. It is also difficult to get object instance masks and their 6D pose information due to occlusion. Therefore realistic datasets are usually small in size and have sparse information. The human labeling process [5] is another potential concern because it is hard to generalize to other systems, and the accuracy of annotating suction grasp poses as ground truth is questionable. More recent training on synthetic dataset approaches can help reduce the cost of data collection but still have limitations regarding the analytical model accuracy, domain gap, fixed vision system, neglecting objects and suction cup gripper dynamics, and insufficient information on cluttered environments. Furthermore, different neural network-based learning methods [4]–[6], [14], [15] for suction cup grippers have been proposed to predict the grasp success using pixel-wise affordance with RGB or depth inputs. These 2D affordance methods rely on images that experience some issues with generalization to unseen objects. Learning on 3D point clouds provides better generalization [16] for novel objects. However, to our knowledge, no study on suction cup grasp success prediction uses object-aware point-wise affordance, which directly takes the 3D point cloud and text prompt as input and generates robust 6D suction grasp poses for object instances. Additionally, the current evaluation metrics [2], [5] for suction grasp prediction precision by comparing the pre-annotated suction grasp ground truth with the prediction result, suffer from the fact that the pre-annotated suction grasp dataset can only include a subset of all possible candidates depending on the sampling number, due to the nature of infinite suction grasp poses existing in every cluttered environments [17]. Together, these limitations demand a universal, accurate, and efficient dataset generation process and network architecture and evaluation metrics, which can serve as a benchmark for developing suction grasp policies.

In this paper, we propose ***Sim-Suction***, a deep learning-based system that uses a suction cup gripper to pick up novel objects from cluttered environments (Fig. 1). It consists of two components: (1) ***Sim-Suction-Dataset***: a large-scale synthetic dataset for cluttered environments, and (2) ***Sim-Suction-Pointnet***: an object-aware point-wise affordance policy that predicts the most robust suction grasp pose for the target object. The primary contributions of our work include:

- A large-scale synthetic dataset for cluttered environments that include RGB images, depth information, single-viewed point clouds, multi-viewed point clouds, object instance segmentation masks, 6D object poses, 2D object bounding boxes, 6D object bounding boxes, camera matrices, 6D suction grasp poses, and 3D suction grasp score maps.
- A suction grasp candidate evaluation process that integrates an analytical model and simulation, assessing object collision, seal formation, suction cup gripper dynamics, and multi-body rigid dynamics in cluttered environments.
- Physical robot experiments validating the Sim-Suction analytical suction model and comparing it with the DexNet model.
- A novel point-wise affordance network that trains on point clouds and annotated 3D suction grasp score map, which outputs point-wise suction grasp success probability.
- An online evaluation metric capable of assessing the precision of suction grasp predictions across different

benchmarks.
- A thorough ablation study examining the effectiveness of the *Sim-Suction-Dataset* diversity and *Sim-Suction-Pointnet* architecture.
- Simulation and physical robot experiments quantifying the *Sim-Suction-Pointnet* suction grasp success rate without prior knowledge of objects in cluttered environments.

II. RELATED WORK

A. Object Affordance

Object affordance was first introduced by Gibson *et al.* [18], which refers to the ability of an agent to perform actions in a given environment. In the suction grasping community, the vast majority of studies use object suction affordance [4], [5], [6], [14], [15] to indicate the most likely part of objects to make the suction grasp successful based on the unique mechanism and shape of a suction cup. Affordance learning is a variant of the segmentation method, which learns from the suction affordance scores and can adapt to novel objects. Research on the suction affordance learning framework mainly focuses on the FCN-based pixel-wise affordance [19] that only uses RGB-D images or depth to infer the suction grasp success probability at each pixel. UMPNet [20] proposes an image-based policy network that infers closed-loop action sequences for manipulating articulated objects. However, image-based affordance often faces generalization challenges when encountering unseen images. Recent studies have started to explore one-shot [21], [22] or zero-shot [23] image-based affordance to address generalization issues in object grasping, but these methods necessitate extensive datasets and resources for training. With the development of PointNet [24] and PointNet++ [25], the feature extractor can extract the 3D features directly from raw point cloud inputs that have good performance in unseen objects. A number of studies [16], [26], [27], [28], [29], [30] begin to learn point-wise graspable affordance for parallel jaw grippers, but no research has been found that investigate point-wise suction affordance.

B. Semantic Segmentation

Semantic segmentation is an essential component in various robotic applications, particularly object grasping. Several studies have used segmentation masks to facilitate grasp planning [2], [31]–[35]. By providing a detailed representation of object boundaries and spatial relationships, semantic segmentation enables robots to understand the shape, size, and category of objects. This information is essential for calculating feasible grasp points and optimizing grasp strategies. Mask R-CNN [36] has been widely used in robotics applications, such as object grasping, due to its ability to provide precise object localization and segmentation. However, Mask R-CNN does not generalize to unseen objects in novel categories. Recent ground-breaking work promptable Segment Anything Model (SAM) [37] demonstrates promising out-of-box zero-shot image segmentation capabilities in various scenarios without any retraining and fine-tuning. It requires 2D points or 2D bounding boxes prompts to provide instance segmentation. The Grounding DINO model [38] proposes open-set image object detector which incorporates a language model to enhance concept understanding, resulting in more effective object detection for unseen objects. Our work fuses the advantages of SAM [37], Grounding DINO [38], and point-wise affordance to propose a suction-grasping policy to use text prompt input to guide the picking-up task by predicting robust 6D suction grasp poses for objects of interest in cluttered environments.

C. Suction Grasp Dataset

Zeng *et al.* [5] proposed a manually labeled suction grasp dataset from cluttered real-world environments. It requires humans with experience to annotate each pixel in the RGB-D images with a binary value representing suctionable and non-suctionable areas. One major drawback of this dataset is that the dataset size is relatively small and include limited object information. Moreover, the empirical suctionable area labeling process is tedious and introduces potential errors. SuctionNet-1billion [6] addresses this issue by proposing a real-world suction grasp dataset and using the analytical model to annotate RGB-D images captured from two popular cameras. However, the time cost to generate a rich real-world dataset remains a considerable limitation. It is also hard to generalize to different environments and vision settings. Dexnet 3.0 [4] instead generates a synthetic suction grasp dataset and uses an analytical model to annotate on singulated object depth images, which do not contain any information about the cluttered environments. Jiang *et al.* [14] turns to generate a synthetic dataset in cluttered environments. However, it fails to consider the domain gap and provides a limited annotation method by analyzing primitive shapes only. Shao *et al.* [39] proposes a suction grasp dataset used for self-supervised learning in cluttered environments but only contains cylinders of the same size, which is restricted to specific applications.

D. Grasp Candidates Sampling

The grasp candidates sampling process refers to randomly sampling the configurations of the end-effector on the target object to generate a large number of possible grasp configurations, for which the grasp 6D grasp poses cannot be calculated directly. Most research in suction grasp sampling has been carried out to sample suction grasp candidates in point cloud space [4], [6], with a little focus on sampling suction grasp candidates on 2D images [40]. Our main focus is to look at grasp sampling in point cloud space since 6D suction grasp poses are highly dependent on the geometry of the objects. The grasp candidates sampling process in point cloud space can be classified into object-agnostic sampling and object-aware sampling. The object-agnostic sampling process does not need individual object information. It treats multiple objects in a cluttered environment as a single unified object. Object-agnostic sampling algorithms [40] perform a search on the entire point cloud space, which is slow and inaccurate in cluttered environments. The object-aware sampling process can solve the above issues by using the complete information in the entire point cloud and help with further evaluation. In this work, we use our dataset's instance segmentation

mask and object 6D poses to create an object-aware sampling scheme, where each suction grasp candidate is associated with the relevant object information.

E. Suction Grasp Candidates Evaluation

For evaluating suction cup grasp candidates, it is challenging to annotate good and bad suction grasps. Mahler*et al.* [4] first propose a compliant suction contact model which uses a spring system to evaluate the seal formation on the contact surface and quasi-static physics to evaluate the ability to resist external wrenches for the singulated object. Cao *et al.* [6] extended the work by simplifying the quasi-static spring system for checking seal formation and resisting external wrenches on the sigulated object, and performed collision checking in cluttered environments. Zhang *et al.* [15] adopts a similar model to evaluate suction grasp candidates on singulated object but fails to address constraints in cluttered environments. Jiang *et al.* [14] uses a convolution-based method to calculate the suctionable area, assuming the suctionable surface is flat and large enough. The authors evaluate the ability to resist external wrenches by only calculating the normalized distance between the suction location and the center of the suctionable area of each object. Overall, the current suction quasi-static physics model has limitations in cluttered environments because it cannot comprehensively analyze the entire cluttered environment and that can lead to false results, especially when the suction cup gripper tries to grab objects from the bottom of the heap when other objects are stacked on them. It also fails to analyze whether the contact is kept established during the suction cup gripper movement. The current seal formation evaluations only examine the contact surface of a singulated object. However, in cluttered environments, a suction cup may have contact with multiple objects or the ground plane when the suction location is on object edge. The studies presented thus far demand the need for an accurate suction grasp candidates evaluation scheme in cluttered environments. This paper makes an essential contribution to suction grasping in cluttered environments by focusing on the entire suction grasp process, including suction cup gripper and object dynamics, instead of only analyzing the quality of established contacts for singulated object. We combine the analytical model with physics simulations to provide an accurate suction grasp evaluation process, which also serves as an online evaluation metric to calculate the prediction accuracy.

III. PROBLEM STATEMENT

A. Overview

We denote the unstructured environment initial state as $\mathcal{X} = (\mathcal{O}, \mathcal{P}, \mathcal{C}, \mathrm{d})$. Given a single view RGB-D image and registered point cloud $\mathcal{P}$ of an unstructured environment consisting of a set of objects $\mathcal{O}$ captured from a depth camera with known camera matrix $\mathcal{C}$, our goal is to enable the robot to use a vacuum suction cup gripper with suction pad diameter d to pick up object $\mathcal{O}_i$ from $\mathcal{O}$ by selecting the most robust suction cup grasp pose $\mathrm{S}=(\mathrm{R},\mathrm{T}) \in \mathrm{SE}(3)$ from all possible suction pose candidates $\mathbb{S}$, where R represents the suction cup approaching direction and T represents the location of the center of the suction pad.

The binary success measurement of picking up $\mathcal{O}_i$ from $\mathcal{O}$ depends on the state of $\mathcal{O}_i$, each object state specifies the object geometry, the center of mass, 6-DoF pose, friction coefficient, and interactions with its surroundings. Therefore, we need to consider those constraints to predict a successful suction grasp. Data-driven approaches showed their capability to handle these physics constraints by either using human labeling or analytical models. However, limitations still exist regarding the data size, the precision of the analytical model, the ability to generalize to customized objects, and the performance in cluttered contexts. To address the aforementioned limitations, we first propose a method to autonomously create and annotate a large-scale synthetic dataset for cluttered environments using the Omniverse Isaac Sim simulator [41]. This results in the creation of a benchmark dataset called *Sim-Suction-Dataset*. Subsequently, we train the dataset with affordance inference networks, *Sim-Suction-Pointnet*, to generate a point-wise suction grasp affordance map. This map is then combined with a task-oriented semantic segmentation mask to generate the grasping policy and refine 6D suction poses for target objects.

B. Assumptions

We make the following assumptions when developing the *Sim-Suction*:

- Objects $\mathcal{O}$ are rigid bodies made of non-porous material with mass, inertia, velocity, and friction and can interface with static or moving rigid bodies in the unstructured environment;
- The suction gripper can be simulated with a spring-mass model and can create a 6-DoF joint with the object of interest.
- The depth camera has known intrinsic matrix $\mathcal{C}$.

IV. SIM-SUCTION-DATASET

Previous suction grasp datasets mainly focused on the labor-intensive collection and labeling processes, which limit dataset size and ground truth precision. Although analytical models like Dexnet-3.0 can reduce some errors introduced by human labeling, they can only be applied to singulated objects and fail to consider dynamic interactions in unstructured environments. Additionally, real-world datasets do not transfer well to different scenarios due to fixed lighting conditions, camera systems, and difficulties in adding new objects. To overcome these issues, we present methods for generating a large-scale dataset through physics simulation, resulting in the *Sim-Suction-Dataset*. This dataset is the first large-scale synthetic suction grasp dataset in cluttered environments that combines analytic models and dynamic interactions. Our dataset (see Table. I) includes 1550 objects from 137 categories. The objects come from ShapeNet [42], YCB objects [43], NVIDIA Omniverse Assets [44], and Adversarial Objects in Dex-Net [45]. The objects can be categorized into three difficulty levels (Fig. 2): Level 1 includes prismatic and circular solids, Level 2 includes objects with varied geometry, and Level

3 includes objects with adversarial geometry and material properties. The resulting mass and difficulty level distribution are shown in Fig. 2. (d) and (e). The difficulty levels of the objects are determined based on the complexity of their triangle mesh. We use the material density of each object to calculate the object's mass. We select a 1.5 cm radius bellows suction cup and a 2.5 cm radius bellows suction cup for our suction cup gripper. We propose a pipeline to study the grasp correlation of each object in the entire cluttered environment to automate the generation of accurately labeled data. This is achieved by combining sampling-based approaches, analytical model analysis, domain randomization, and dynamic physics simulations. We use Nvidia Omniverse Isaac Sim Simulator [41] with the built-in PhysX engine [46] as a toolkit to simulate rigid body dynamics and annotate 6D suction grasp poses(Fig. 4). Each grasp has a corresponding 6D grasp pose, gripper dimension, object difficulty level, object physics information, RGB-D image, camera matrix, binary success label, and object segmentation point cloud associated with it. This comprehensive dataset and auto-labeling pipeline are intended to serve as a synthetic benchmark and reference for suction grasping research. Notably, the dataset generation pipeline itself could serve as a foundational model-based strategy. This opens doors for researchers to sample suction grasps without a dedicated trained model. Our dataset provides a common dataset and approach for comparison and evaluation of suction grasping algorithms across different gripper sizes and can also be used to develop point-wise affordance grasping policies (Section V-B).

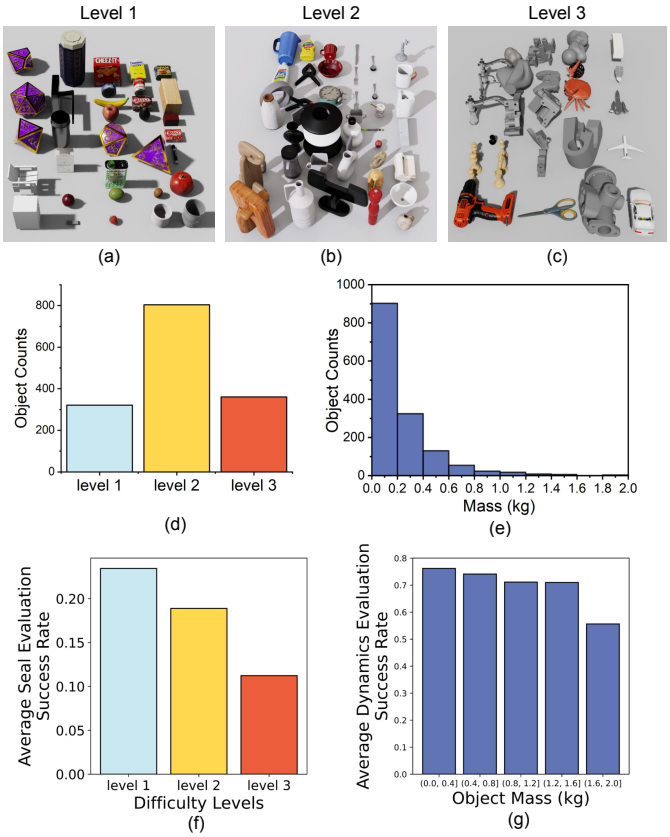


Fig. 2. (a) - (c) Examples of objects with varying difficulty levels, where Level 1 is the least challenging and Level 3 is the most challenging. (d) - (e) Depiction of object mass and difficulty levels distribution within the Sim-Suction-Dataset. (f) Influence of object difficulty levels on seal evaluation. (g) Effect of object mass on dynamics evaluation.

A. Cluttered Scene Generation

The first step of generating the *Sim-Suction-Dataset* is to create random unstructured scenes that contain objects $\mathcal{O}$. In the real world, humans just arbitrarily dump objects $\mathcal{O}$ on a ground plane to create such scenes. We adopt the same strategy in Isaac Sim Simulator [41]. We define our state distribution ζ_{scene} as a product of the following:

- Total objects number I in one scene: Randomly selected from a range $[1, 20]$.
- Objects $\mathcal{O}$: uniformly randomly sampled with replacement of total size I from 20 selected 3D models from YCB-dataset. Objects $\mathcal{O}$ drop locations are uniformly sampled from the 3D space $[-0.1\ m, -0.1\ m, 0.5\ m] \times [0.1\ m, 0.1\ m, 0.8\ m]$ above the ground plane. Objects $\mathcal{O}$'s orientations are uniformly randomly sampled. Each $\mathcal{O}_i$ is dropped at free fall with mass m_i.
- Coulomb friction coefficient μ: Randomly sampled from the range $[0, 1]$ and used to model tangential forces between contact surfaces.

For the simulator to implement multi-body physics, we use convex decomposition to describe the collision geometry, where several convex shapes approximate the input mesh. We first sample the initial state from the state distribution ζ_{scene} to begin the data generation process. Then we start the dynamics simulation and drop objects one-by-one with a rendering time step to avoid object penetration until all objects on the ground

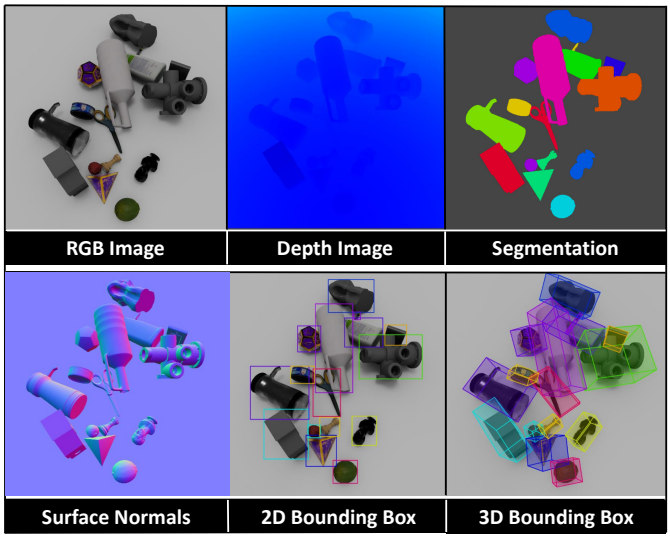


Fig. 3. The Photo-realistic RGB-D images are rendered from a synthetic camera with an intrinsic matrix sampled around the nominal values of a PrimeSense Carmine 1.09 camera. The segmentation mask is generated using the GPU-RayTracing in PhysX engine [46]. The segmented point cloud can be registered from 2D RGB-D images and surface normals with the help of a segmentation mask using camera intrinsic and extrinsic matrices. We also provide the 2D and 3D bounding box labels for each object instance, which can contribute to the object detection and pose estimation community.

TABLE I
COMPARISON OF SUCTION GRASPING DATASETS.

Dataset	Grasp Pose Label (Method)	Objects/ Scene	Camera Type	Total Objects	Total Labels	Modality	Multiple Gripper sizes	Semantic Segmentation	Dynamics Evaluation
SuctionNet [6]	6D (✎)	∼10	Real	88	∼1.1B	RGB-D	No	Yes	No
Dex-Net 3.0 [4]	2D (✎)	1	Sim	1.6K	2.8M	Depth	No	No	No
A. Zeng [5]	2D (👤)	NA	Real	NA	191M	RGB-D	No	No	No
Sim-Suction-Dataset	**6D (✎, ▶)**	**1-20**	**Sim**	**1.5K**	**3.2M**	**RGB-D**	**Yes**	**Yes**	**Yes**

Note: Grasp labels can be generated either manually (👤), using analytical models (✎), or through physics simulation (▶). Dynamics evaluation is denoted as partial when it only evaluates an isolated single object, rather than objects in cluttered environments.

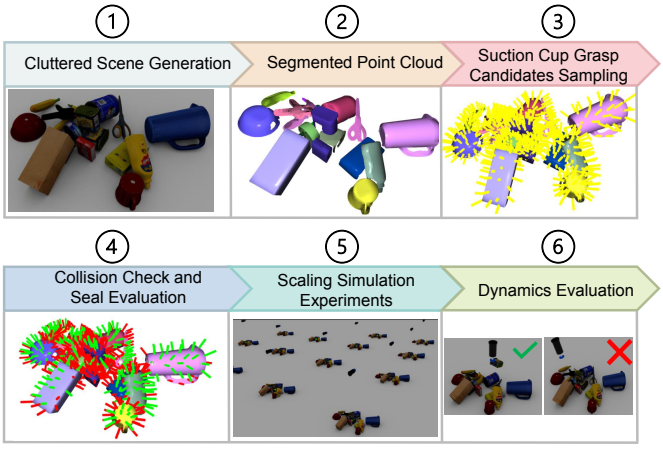


Fig. 4. Sim-Suction-Dataset generation pipeline. ① Objects free fall above the ground plane to create a cluttered scene. ② Create segmented point clouds for each scene from multi-viewed synthetic camera. ③ Sample suction grasp candidates from the object surface. ④ Evaluate each candidate with a combination of the analytical model and PhysX engine to check whether it fails to form a seal or has collisions with surroundings. ⑤-⑥ Use simulation to further evaluate each candidate under realistic dynamics settings.

plane reach their static equilibrium. Finally, we repeat this process to populate the dataset with 500 cluttered scenes.

B. Object-aware Suction Grasp Sampling

Previous sampling methods for cluttered environments use an object-agnostic sampling strategy, where the sampling algorithm searches the point clouds in the entire scene. This method causes a low sampling accuracy and a time-consuming sampling process. It also has poor grasp candidate coverage among different sizes of objects. Additionally, the object-agnostic sampling strategy cannot associate the sampled grasp candidates with object instances and object poses. It considers the point cloud of the cluttered scene as a whole and thus prevents further evaluation. Although the state-of-art semantic segmentation and 6D pose estimation algorithms can play a part in retrieving object information, we question the estimation accuracy as "ground truth". To overcome these limitations, we propose an object-aware sampling strategy, which combines the suction grasp candidates with their object information. Owing to the GPU-RayTracing technology in PhysX Engine, we can easily extract each instance's 6D pose and segmentation information from the cluttered environment and reduce the sampling algorithm searching space to each object's point cloud rather than the entire scene. It is commonly assumed that all possible suction poses are infinite in $\mathrm{SE}(3)$ for a single object. Thus, it is impossible for the sampler to cover all of them. Our suction grasp sampling process aims to find a large set of suction candidates evenly distributed on the object surface by considering the suction pad diameter and the time cost for annotating.

*1) **Cluttered Scene Point Cloud Processing**:* Our suction grasp sampling process relies on the complete geometries of the object $\mathcal{O}$ in the cluttered scene. In order to get a good geometric description of the cluttered environment due to the existence of object occlusions, we use 800 synthetic cameras to register segmentation point clouds in $\mathrm{SE}(3)$ and merge the multi-viewed point clouds into a single point cloud $\mathcal{P}$ representing the cluttered scene, then we calculate the surface normals of each point to get local geometry. We save individual object point clouds and cluttered environment point clouds for each scene.

*2) **Geometry Guided Approach-Based Sampler**:* given the point cloud of each object instances i from a cluttered scene, we aim to find a set of suction grasp candidates as $\mathbb{S}$=($\mathbb{R}$,$\mathbb{T}$) $\in$ SE(3), which describes the pose and orientation of the suction gripper. To form a seal, we want to align the suction gripper approaching vector with the objects' surface normal on a sampled suction point $t \in \mathbb{T}$. We use iterative Farthest Point Sampling (FPS) [47] to choose a set of points $\mathbb{T}$ from each object point cloud. It is considered to have better coverage on the object surface over the random sampling process. For each sampled suction point t on the differentiable object surface, we calculate the corresponding Darboux frame as $\mathrm{R} \in \mathbb{R}$. A Darboux frame is a natural moving frame constructed on a surface to study curves:

$$R(t) = [v_1(t)|v_2(t)|v_3(t)], \tag{1}$$

where $v_1(t) \in \mathcal{N}$ is the normal vector, $v_2(t)$ is the major axis of curvature vector, and $v_3(t)$ is the minor axis of curvature vector. We calculate $v_1(t)$,$v_2(t)$ and $v_3(t)$ by evaluating the Eigenvectors of the 3×3 matrix $N(t)$:

$$N(t) = \sum_{t \in \mathbb{T}} \hat{n}(t)\hat{n}^T(t), \tag{2}$$

where $\hat{n}(t)$ is the normal vector at point t. $[v_3(t), v_2(t), v_1(t)]$ is the Eigenvectors of matrix $N(t)$ in decreasing order. We only align our suction grasp candidates' X-axis (Fig. 5) with $v_1(t)$ to ensure that the suction cup makes full contact with the object's surface. The $v_2(t)$ and $v_3(t)$ of the Darboux frame (tangent and binormal vectors) might not be as crucial for suction cup grippers as they are for parallel-jaw grippers, but they can still provide additional information about the local geometry of the object's surface that is useful in seal evaluation and grasp planning.

Fig. 5. **Left (1.5 cm radius bellows suction cup).** We evaluate the seal performance by casting dense rays along surface normal vectors from the suction cup surface towards the object surface. To evaluate the suction dynamics, we model the suction cup gripper with a 6 degree of freedom joint. We set the suction cup bending angle limit to lock individual axes. We set 20 N force limit for 1.5 cm suction cup and check if the 6D joint can be created and maintained during the manipulator movement. **Right (2.5 cm radius bellows suction cup).** We set the 30 N force limit for 2.5 cm suction cup.

C. Suction Grasp Candidates Evaluation

Robots with suction cup grippers interacting physically with the objects in cluttered environments face inherent uncertainties in how objects will react to suction. Previous methods fail to consider the correlation between the object of interest and its surroundings and only evaluate grasp candidates on the singulated object. We believe such methods decrease the accuracy of ground truth labeling by introducing False Positives. We propose a new suction grasp candidates evaluation pipeline, which evaluates the entire cluttered environments with the support of PhysX Engine to provide accurate annotations. The pipeline has three ordered sub-evaluation sequences with binary-valued metric on each suction candidate S : Collision check $\mathcal{Q}_{collision}(\mathrm{S}) = \{0, 1\}$, Seal formation evaluation $\mathcal{Q}_{seal}(\mathrm{S}) = \{0, 1\}$, and dynamics evaluation $\mathcal{Q}_{dynamics}(\mathrm{S}) = \{0, 1\}$. We define the final metric as a product of the sub-evaluation metrics $\mathcal{Q}(\mathrm{S}) = \mathcal{Q}_{collision}(\mathrm{S}) \times \mathcal{Q}_{seal}(\mathrm{S}) \times \mathcal{Q}_{dynamics}(\mathrm{S})$. More detail about each sub-evaluation system is described next:

*1) **Collision Check and Seal Evaluation**:* Suction cup grippers can lift an object when the pressure difference between the atmosphere and the vacuum is large enough. Intuitively, a suction cup gripper can easily form an airtight seal on a flat surface. However, forming a seal with the suction cup becomes a challenge when dealing with an irregular surface. To address this issue, we took inspiration from a spider's web and model the bellows suction cup with 15 concentric polygons, each with 64 vertices. We perform the collision check in a physics simulator by casting rays along the x-axis from each vertex, as shown in the Fig. 5, to detect the closest object that intersects with a specified ray. We evaluate the suction cup's seal by modeling it with deformable material and as a spring, with a deformable threshold of 10%. In Fig. 2. (f), the negative correlation shows that the candidate seal evaluation passing rate decreases with increasing object geometry complexity. The trend indicates that it is more difficult for a suction cup gripper to form a seal on a complex object surface. Sim-Suction with a 960-vertex suction cup model can be utilized for assessing suction seals on intricate geometries, including uneven surfaces and surfaces with holes or grooves. We provide a comprehensive comparison of corner cases as shown in Fig. 6. In unstructured environments with different difficulty levels of objects, the ground truth labeling process is expected to handle different scenarios and provide an accurate result. However, the previous suction model proposed by Dex-Net and used by others [6], [15] has limitations in dealing with complex geometries and overlapped environments. The DexNet model utilizes the perimeter, flexion, and cone spring connected by eight vertices to assess the seal formation. These vertices are selected on the outer perimeter, resulting in the neglect of any geometry inside the suction cup perimeter. Theoretically, if the suction cup gripper's radius is small enough, and the object is non-porous, there would be no need to be concerned about any geometry inside the suction cup causing air leaks. However, in reality, suction cup grippers are usually larger than the small features commonly found on objects rendering them impossible to ignore. Fig. 6 (a) shows that a false positive when all eight vertices of the DexNet model sit on a flat surface and the spring deformations are within the threshold, but there is a groove under the suction cup gripper. Fig. 6 (b) and (e) show that there are geometries under the suction cup gripper that cause the suction cup to deform and not create proper seal. Fig. 6 (c) shows that the DexNet model resolution is not suitable for rough surfaces. Fig. 6 (d) and (f) show that the DexNet analytical model only takes in the singulated object information. In cluttered environments, DexNet fails to identify neighboring objects, and its collision check performed in DexNet cannot handle these scenarios without object segmentation information. To better quantify our suction model performance, we design and print a 1:1 digital-twin testing board (Fig. 6. (g)) consisting of various challenging features such as holes, rough surfaces, and complex geometries. We perform validation experiments with the candidates $\mathcal{Q}_{seal}(\mathrm{S}) = 1$. The results from Table. II show that Sim-Suction suction Model provides more accurate annotations compared to the DexNet suction model. The failure cases in the Sim-Suction experiments are primarily caused by the Apriltag [48] precision error and imperfections in 3D printing.

*2) **Dynamics Evaluation**:* Previous methods primarily focus on analyzing the external wrench acting on a singulated object. In cluttered environments where the object of interest is

TABLE II
COMPARISON OF SUCTION GRIPPER ANALYTICAL MODELS ON TESTING BOARD

Model	Total Grasps	Successful Grasps	Success Rate
DexNet	136	83	61.03%
Sim-Suction	160	155	96.88%

This table compares the performance of the DexNet and Sim-Suction suction gripper analytical models on a specially designed testing board consisting of various challenging features. The success rate indicates the percentage of successful grasps out of the total attempted grasps for each model.

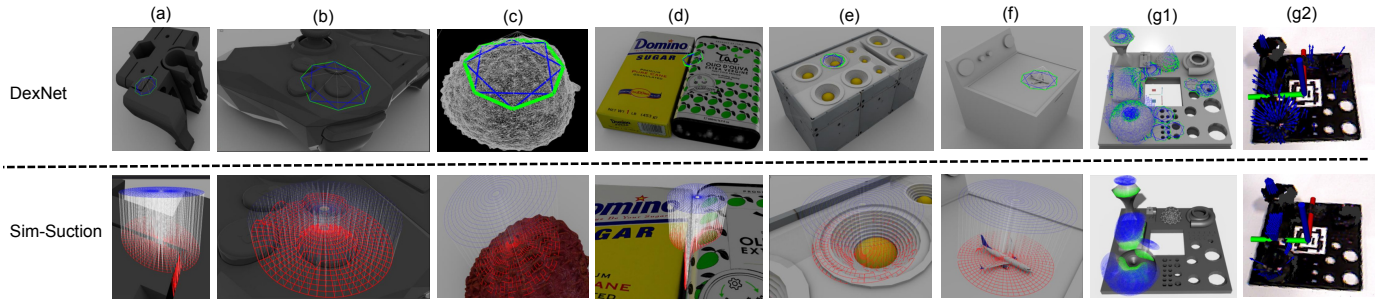


Fig. 6. Comparison for 1.5 cm suction cup gripper model. (a) Surface with grooves and holes. (b) Protruding parts. (c) Rough surface. (d) Objects next to each other. (e) Concave surfaces. (f) Overlapped objects. (g1) Customized complex testing board. (g2) 3D printed testing board.

at the bottom of a pile, the suction cup gripper must resist not only the wrench due to the object of interest's gravity but also that of the objects above it. Moreover, these methods neglect the dynamics between the suction gripper and the objects. Eppner *et al.* [49] demonstrated that simulations considering the entire grasp process, including dynamics, yield more information about grasp success compared to analytical quality evaluations that only measure static contact quality. To achieve a more accurate $\mathcal{Q}(\mathrm{S})$, we employ a 1.5 cm radius suction cup mounted on a 7-DoF UR10 robot to simulate the dynamics $\mathcal{Q}_{dynamics}$ using the GPU-enabled Isaac Sim simulator. This simulator utilizes reduced coordinate articulations with Temporal Gauss Seidel (TGS) [50] to compute the future states of objects and the suction cup. We use Riemannian Motion Policy (RMP) [51] to control the UR10 manipulator to reach suction pose configuration S=(R,T). We model the suction cup gripper as a D6 joint, which represents a 6-degrees-of-freedom constraint defining the relationship between the gripper and the object being grasped. The joint constrains the relative position and orientation of the suction cup and the object, permitting them to function as a single entity. The Temporal Gauss-Seidel (TGS) method is an iterative solver used in physics simulations to compute the future states of objects and the suction cup gripper by solving constraint-based systems efficiently. We calculate $\mathcal{Q}_{dynamics}(\mathrm{S})$ by determining whether the suction cup can create and maintain a 6D joint with suction candidate S between the object surface and the suction cup surface in a dynamic environment. This process considers various parameters, including force limit, torque limit, friction coefficient, object mass, bellows suction cup maximum bending angle, stiffness, and damping rate. These parameters are essential for TGS to accurately simulate the dynamics and performance of the suction cup gripper modeled as a D6 joint in diverse scenarios, assessing its effectiveness in grasping objects in cluttered environments.We set the force limit and torque limit using data obtained from the silicone 1.5 cm suction cup gripper. These limits represent the maximum force and torque that the gripper can hold before breaking the constraint, affecting the stability and strength of the D6 joint. TGS uses these parameters to decide if the joint can withstand the forces acting upon it during the simulation. The bend angle defines the maximum angle that the suction cup gripper can bend when a load is applied. It helps TGS simulate the deformation of the suction cup when subjected to forces and torques, ensuring that the gripper can maintain a seal with the object surface. The bending stiffness represents the resistance of the suction cup gripper to deformation. TGS uses this parameter to compute the forces and torques required to maintain the shape of the suction cup and ensure proper contact with the object surface. The bend damping parameter helps TGS simulate the energy dissipation in the suction cup gripper during deformation. It contributes to the overall stability and realism of the simulation, particularly when the gripper experiences dynamic forces and torques. By accounting for these parameters, the TGS solver can accurately simulate the dynamics and performance of the suction cup gripper modeled as a D6 joint in various scenarios. We conduct the GPU-based multi-task simulation by trying to lift the objects $\mathcal{O}$ with suction cup configurations $\mathbb{S}$ after passing collision and seal evaluation. In Fig. 2. (g), the negative correlation shows that the candidate dynamics simulation evaluation passing rate decreases with increasing object mass. The trend indicates that it is more difficult for a suction cup gripper to lift a heavy object from various directions.

V. Suction Grasp Estimation Network

In this section, we describe our object-aware suction grasp pose estimation network in detail, *Sim-Suction-Pointnet*.

A. Dataset Preprocessing

The point-wise affordance networks require a binary pixel mask. Given suction grasp candidates $\mathbb{S} = (\mathbb{R}, \mathbb{T})$ with $\mathcal{Q}(\mathbb{S}) = 1$ for each cluttered environment, we use a ball query algorithm on a complete point cloud $\mathcal{P}$ to find all points $\mathcal{P}_t$ that are within a radius of $1.5\ cm$ to the query point $\mathrm{T} \in \mathbb{T}$ to represent the contact points between the suction cup and the object surface, and annotate point set $\mathcal{P}_t$ with binary score 1, and the complement point set $\mathcal{P'}_t$ with binary score 0. We use the annotated score for each point as the binary point mask for *Sim-Suction-Pointnet*.

B. Sim-Suction-Pointnet Framework

*1) **Affordance Network**:* The framework for *Sim-Suction-Pointnet* showed in Fig. 7. The *Sim-Suction-Pointnet* is to learn object-aware suction affordance grasping policy in 3D space. We use PointNet++ as our backbone network. The PointNet++ takes raw point clouds $\mathcal{P}$ into the sampling layer,

which uses the farthest point sampling (FPS) to choose and normalize a subset of points containing N points with d-dim coordinates. The normals layer takes an $N \times d$ matrix as input and outputs an $N \times (d + M)$ matrix, where M is additional point feature channel. We use surface normals for M suggested in [25] as it can increase semantic segmentation performance. We modify the PointNet++ network with the parameters are shown:
$SA(5120, 0.02, [128, 128, 256]) \rightarrow$
$SA(1024, 0.08, [256, 256, 512]) \rightarrow$
$SA(256, 0.2, [512, 512, 1024]) \rightarrow FP(1024, 1024) \rightarrow$
$FP(512, 512) \rightarrow FP(256, 256, 256)$,
where $SA(K, r, [l_1, ..., l_d])$ is a set abstraction (SA) level with K local regions of ball radius r using PointNet of d fully connected layers with width $l_i (i = 1, .., d)$; $FP(l_1, ...l_d)$ is a feature propagation (FP) level with d fully connected layers. The decoder of PointNet++ is to turn the group features into point-wise features. The PointNet++ loss, $\mathcal{L}_{score}$, is based on MSE loss:

$$\mathcal{L}_{score} = \frac{1}{N} \sum_{p \in \mathcal{P}} (\mathcal{Q}_p - \hat{\mathcal{Q}}_p)^2, \tag{3}$$

where $\mathcal{Q}_p$ is the ground truth point score of point p, and $\hat{\mathcal{Q}}_p$ is the predicted probability score of point p. We trained the network on complete point clouds from 500 cluttered scenes with a learning rate of 0.001. To augment the dataset during training, we uniformly randomly select points as centroids and choose $10,000$ nearest points around each centroid. We further augment the dataset by jigging with small rotation angles and scaling to different sizes resulting to $813,451$ point clouds.

*2) **Object Detection and Segmentation Mask** :* In complex settings, the depth sensors are noisy. The segmentation models trained on RGB have been shown to produce accurate semantic masks [52]. For Sim-Suction-Pointnet, we use the synergy of the point cloud for generating point-wise affordance and RGB image for generating object semantic mask. We add zero-shot Grounding DINO [38] as an object detector fine-tuned with *Sim-Suction-Dataset* that takes text prompt as input to generate object bounding boxes, and zero-shot Segment Anything (SAM) [37] that takes bounding boxes as prompt to generate semantic segmentation mask. Zero-shot object detection and segmentation mask methods offer significant benefits when dealing with the challenges of recognizing and segmenting objects from diverse categories without any prior training examples that can transfer knowledge to unseen classes.

*3) **ScoreNet**:* We integrate a multilayer perceptron (MLP) and output head into our approach to regress and smooth the extracted point-wise features into $N \times 1$ suction probability scores. Utilizing the instance segmentation masks from SAM, we identify object boundaries and filter out suction poses that may result in collisions with other objects in the scene by analyzing the segmentation masks of neighboring objects. We calculate the distance between the centroid or bounding box of the target object and those of other objects in the scene, determining safety margins around each object. These safety margins represent the minimum distance the suction cup should maintain from the object's boundary to avoid collisions. We employ the Darboux Frame to generate 6D suction grasp poses. If a suction pose candidate is found to overlap with the safety margin of any neighboring object, we remove that candidate from the list of potential suction poses, ensuring that the remaining suction poses are collision-free with respect to neighboring objects. Finally, we rank the refined instance suction pose candidates based on their updated suction grasp affordance scores.

VI. EXPERIMENTS

In this section, we first utilize an online evaluation system to explore the effects of dataset diversity and point cloud type, comparing the results with baselines on both similar and novel datasets. Subsequently, we perform an ablation study with real robot experiments to investigate the impact of the segmentation mask, contrasting our findings with baseline approaches. Lastly, we conduct extensive real robot experiments to assess the suction grasp success rate of our *Sim-Suction* policy when retrieving novel objects from a variety of cluttered environments and compare it with state-of-the-art methods.

A. Online Ablation Study

Previous methods employ an offline evaluation system to calculate the Average Precision (AP) by comparing the inferred affordance score with the pre-annotated ground truth from the dataset. However, pre-annotated ground truth cannot cover all possible suction poses that exist, leading to inaccurate AP results. To address this issue, we adopt the online evaluation system from Section IV.D. Given a set of 6D suction configurations $\mathbb{S} = (\mathbb{R}, \mathbb{T})$ and the corresponding confidence scores after inference, we consider a suction pose $\mathrm{S} \in \mathbb{S}$ as a true positive if $Q(\mathrm{S}) = 1$, where $\mathcal{Q} = \mathcal{Q}seal \times \mathcal{Q}collision \times \mathcal{Q}_{dynamics}$. We conduct all performance evaluation experiments on an NVIDIA RTX 3080 Ti GPU.

Our baseline method uses the single-viewed point cloud to predict the affordance score by estimating the variance of the surface normals on each point with its nearby neighbors using ball query with a radius of $1.5\ cm$. The baseline method aims to calculate the object surface flatness around that point, where high variance means low flatness. We use an instance segmentation mask to remove the ground plane because it is not our region of interest and has the highest flatness score. Sim-Suction-Pixelnet method uses DeepLabV3+ as backbone and trained with Sim-Suction-Dataset that takes RGB-D images as input and outputs a pixel-wise affordance map. We use the same 6D pose layer from *Sim-Suction* to process all affordance scores and output the 6D suction grasp poses. Table III presents the performance of various networks under different training sizes and test conditions, comparing their Average Precision (AP) on both similar and novel datasets. Similar objects refer to objects that share common characteristics with those in the training dataset but with different scales. Novel objects are objects that are introduced during the testing phase and are not part of the training dataset. The results demonstrate the importance of dataset diversity, point cloud type, and point-wise learning in training the models for improved grasp prediction.

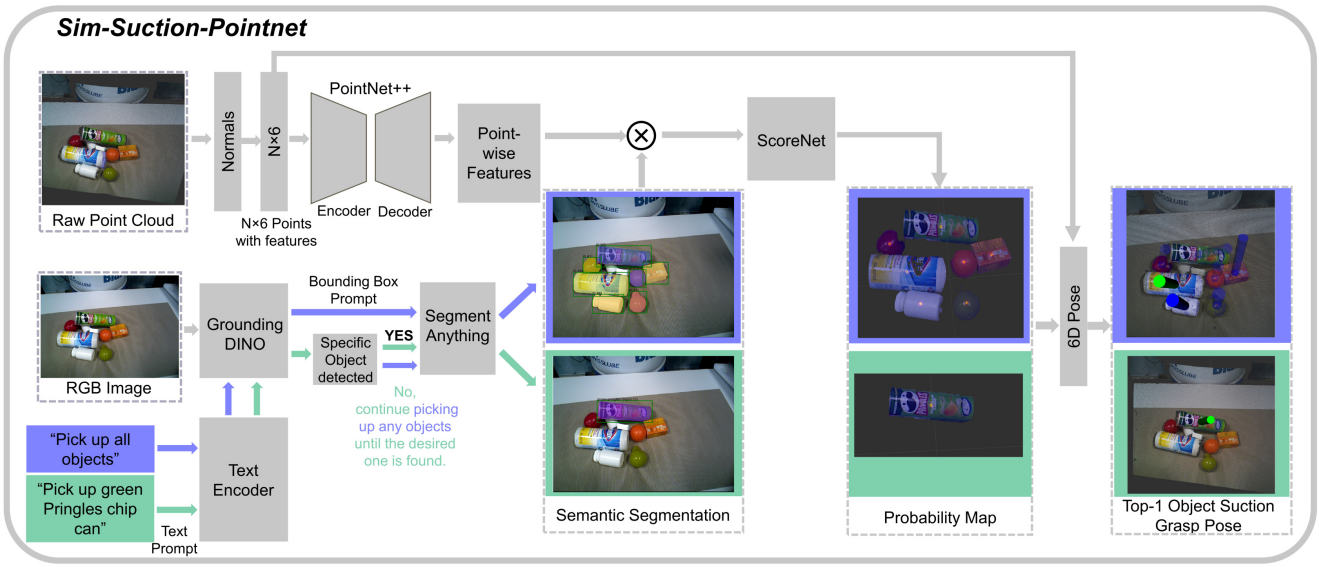


Fig. 7. The Sim-Suction 6D suction grasp pose policy. The green marker represents the 6D grasp pose for the object instance with the highest confidence score. The transparency of the blue markers indicates the confidence score, with higher transparency implying lower confidence and vice versa.

*1) **Dataset Diversity**:* To demonstrate the importance of a large-scale dataset, we evaluate the performance of our method on both a similar dataset and a newly generated novel dataset, which includes 100 unique objects. The *Sim-Suction-PointNet* performance increases with the increase in dataset diversity.

*2) **Effect of Point Cloud Type**:* To illustrate the rationale for training on complete point clouds merged by multi-view camera frames for *Sim-Suction-PointNet*, we evaluate the performance and compare it to *Sim-Suction-PointNet* trained with a single-viewed point cloud. The results show that *Sim-Suction-PointNet* achieves slightly better performance across all dataset diversities, even when inference is performed using a single-view camera. One possible reason is that multi-view merged point clouds provide a more complete and detailed representation of the object, capturing its various features and geometries from multiple perspectives. This richer representation enables the model to learn more robust and generalized features during training, leading to better performance during inference.

*3) **Effect of Point-Wise Learning**:* Our *Sim-Suction-PointNet*, trained with point clouds, demonstrates better performance on novel objects compared to Sim-Suction-PixelNet, which is trained with RGB-D images. One possible reason is that point clouds directly represent the 3D structure of the scene, providing precise geometric information about the objects. This information enables the model to better understand the shape and size of the objects, which in turn helps it learn more effective grasp affordances for novel objects. Point cloud representations are more invariant to viewpoint and scale changes compared to RGB-D images. This allows the model to generalize better across different object orientations, sizes, and camera viewpoints, leading to improved performance on novel objects.

B. Real Robot Experiments Setup

To further evaluate the *Sim-Suction* performance in the real world and address the domain gap problem, we perform experiments with a Fetch mobile manipulation platform equipped with a Primesense Carmine 1.09 head camera and a modular end-effector system [3] with interchangeable 1.5cm radius suction cups with multi-bellow designs rated for 1.3kg payload (Fig. 9). The inference and grasping planning algorithms run on a remote laptop with an NVIDIA GeForce 3070Ti GPU. The Fetch robot and the vacuum pump control module communicate with the remote laptop via ROS nodes. To initiate the experiments, the Fetch robot approaches the workbench and positions itself to observe the tabletop. Once in its initial position, the robot's base remains static throughout the operation, as base movement is not required for arm movement and grasp planning in our setup. While the base is fixed, the robot's torso is capable of vertical movement to adjust its viewing angle and arm height as needed, which is considered part of the motion planning. As a result, the camera height may vary across trials due to the torso adjustments, creating an arbitrary viewpoint for each experiment and testing the model's adaptability. Given the limitations of our camera setup in capturing fine object details, we made selective decisions about the objects included in the real-world experiments. Specifically, Level 3 objects with intricate details were excluded, as our manipulator's camera resolution was insufficient to accurately capture their nuances, affecting grasp prediction performance. Our focus was primarily on Level 1 and Level 2 objects, as these categories represent most objects commonly encountered. We selected 60 novel objects for our experiments, which the policy had no prior knowledge of. The objects were split into two difficulty levels shown in Fig. 9: Level 1: 20 objects with only primitive shapes, and Level 2: 40 objects with varied geometries. For Level 1, since it has fewer objects, we dump all 20 objects onto the table to create a confined environment. For Level 2, we place the 40 objects

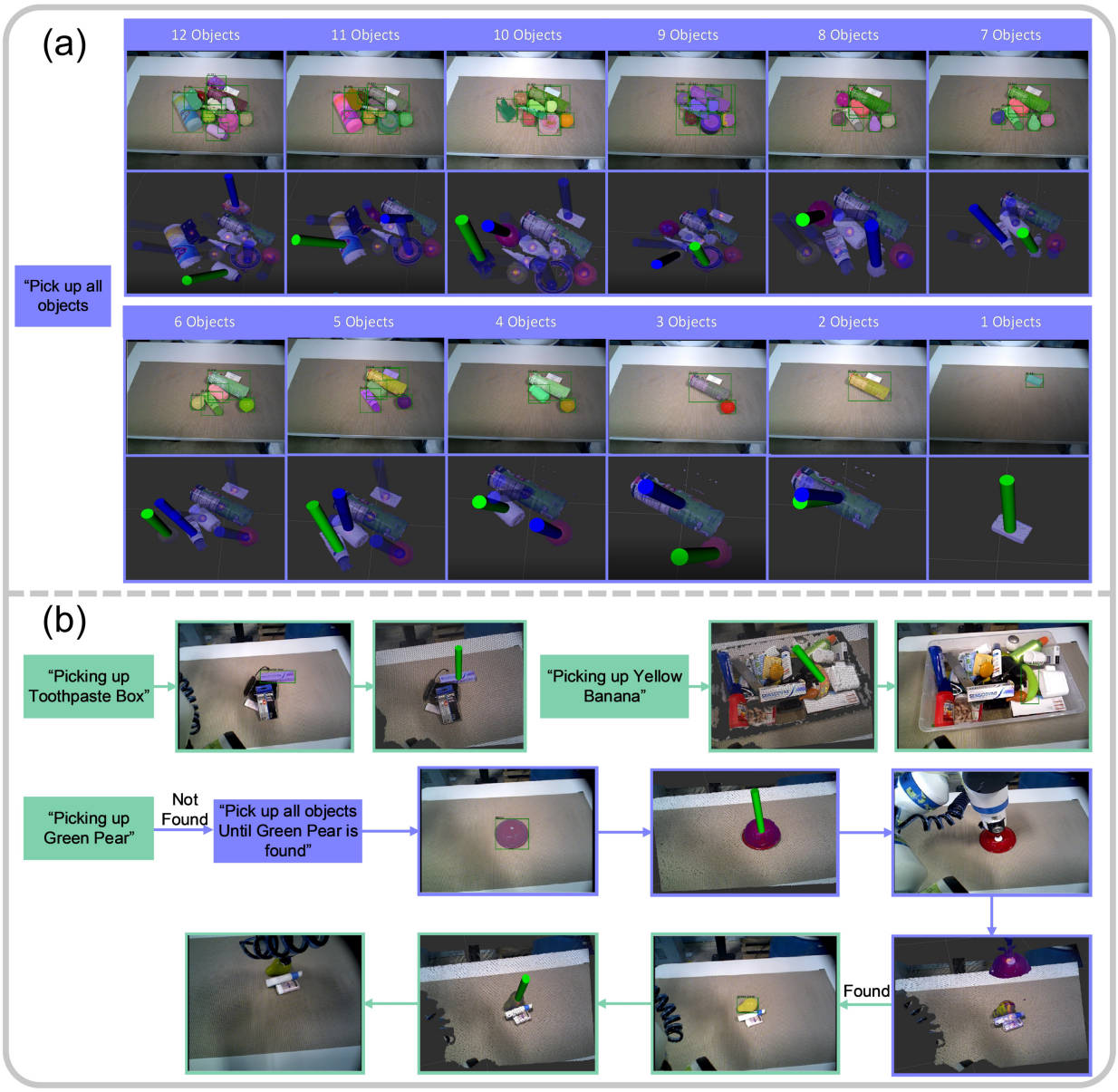


Fig. 8. The *Sim-Suction* policy task sequence examples. The policy demonstrates robust grasping reliability in real-world scenarios. The figure displays the policy applied in two tasks: (a) "pick up all objects", where the robot continuously attempts grasps until the table surface is clear, and (b) "pick up a specific object", where the policy focuses on grasping a target object based on the text prompt input.

TABLE III
ONLINE ABLATION STUDY OF NETWORKS FOR DIFFERENT TRAINING SIZES AND TEST CONDITIONS

Training Size	Network	Test Similar				Test Novel			
		Top-1	Top-1%	Top-5%	Top-10%	Top-1	Top-1%	Top-5%	Top-10%
20 objects, 100 scenes	Baseline	66.34	64.96	60.44	52.66	65.12	64.25	59.76	51.43
	Sim-Suction-Pixelnet	88.04	85.85	79.87	74.77	77.01	74.27	69.91	65.36
	Sim-Suction-Pointnet (SV-PCL)	84.81	79.95	75.98	68.61	81.41	77.84	76.2	66.56
	Sim-Suction-Pointnet (MV-PCL)	85.72	81.86	77.46	71.92	83.43	80.54	77.87	70.36
20 objects, 500 scenes	*Sim-Suction-Pointnet (SV-PCL)*	86.63	83.77	78.94	75.23	85.45	83.24	79.54	74.16
	Sim-Suction-Pointnet (MV-PCL)	87.54	85.68	80.42	78.54	87.47	85.94	81.21	77.96
1550 objects, 500 scenes	*Sim-Suction-Pointnet (SV-PCL)*	88.45	87.59	81.9	81.85	89.49	88.64	82.88	81.76
	Sim-Suction-Pointnet (MV-PCL)	**89.36**	**89.5**	**83.38**	**81.16**	**91.51**	**91.34**	**84.55**	**82.56**

Abbreviations: SV-PCL refers to a model trained on single-view point clouds, while MV-PCL refers to a model trained on multi-view merged point clouds. Top-1, Top-1%, Top-5%, and Top-10% represent the performance metrics for different confidence percentiles.

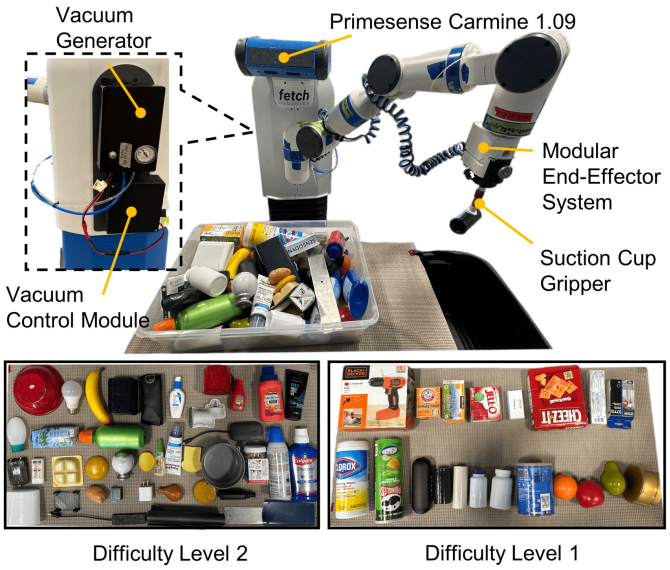


Fig. 9. **(Top)** The experimental setup with a Fetch robot equipped with the Modular End-Effector System [3]. **(Bottom)** We choose 60 household items, with 20 objects in Level 1 (primitive shapes) and 40 objects in Level 2 (varied geometries). These objects are considered novel to the *Sim-Suction-Pointnet* policy, as it has no prior knowledge of them. The objects feature a range of challenging characteristics, such as complex geometries, irregular shapes, and varied surface textures, making the task more difficult.

in a bin. For the cluttered mix, we put all 60 objects in the bin. This experimental setup further tests the performance of the *Sim-Suction* policy in handling objects with different shapes and difficulty levels under various environmental conditions.

C. Experimental Results

We employ a strict reliability metric to evaluate the *Sim-Suction* performance on grasping the selected objects: the ratio of the number of successful grasps to the total number of attempts. This metric is stringent as it accounts for every individual attempt, without aggregating any subsequent attempts even if the item is ultimately grasped. This assessment highlights the policy's ability to rapidly and accurately identify suitable grasping points on novel objects, underscoring the significance of robust performance in each grasp attempt. The ultimate test of *Sim-Suction* is to execute the policy in the real world and deal with domain gap. We want to show our *Sim-Suction* policy trained on large-scaled synthetic point cloud data can transfer well to reality and achieve robust grasp reliability. Fig. 8 shows the Sim-Suction policy on example tasks. For the "pick up all objects" task (Fig.8. (a)), the robot performs continuous grasp attempts until no objects remain on the table surface. This task is not sensitive to the text prompt input. For the "pick up a specific object" task (Fig.8. (b)), if the object of interest is found, the policy executes the pick-up; otherwise, the policy will first carry out the "pick up all objects" task to search for the object of interest. If found, the policy will pick up the target object and complete the task. This task is highly sensitive to the object description provided by humans as a text prompt. We primarily concentrate on the "pick up all objects" test for several key reasons. First, it enables a thorough assessment of the *Sim-Suction* policy by challenging its adaptability and versatility across a wide range of objects with different shapes, sizes, and geometries. Second, focusing on this task helps evaluate the policy's robustness in terms of continuous performance, providing insights into the model's reliability and efficiency in real-world settings. Moreover, the "pick up all objects" task is less sensitive to the text prompt, which allows us to focus on the core aspects of the grasping policy. We initially conduct an ablation study on a subset of testing objects to investigate the impact of segmentation masks on improving the success rate for the "pick up all objects" task and compare it to the baselines. Then, we conduct comprehensive experiments to pick up all 60 objects via a series of experiments to compare the *Sim-Suction* performance against the current state-of-the-art method DexNet 4.0 Suction (FC-GQ-CNN-4.0-SUCTION). We utilize a robot equipped with the MoveIt! motion planning framework to execute the suction grasp with the highest confidence score, which is represented by a green marker in Fig.8. (a). If the motion planning framework fails to find a valid solution, the policy proceeds with the next-best suction grasp, indicated by a solid blue marker in Fig.8. If the motion planning framework continues to fail in finding a valid solution, the policy will proceed with the subsequent suction grasp options based on their confidence scores, in descending order.

*1) **Ablation Study (picking up all objects)**:* We conduct ablation experiments on a subset of test objects, which includes 6 objects from level 1 and 8 objects from level 2. We compare *Sim-Suction-Pointnet (Mask)* with *Sim-Suction-Pointnet (No Mask)* and the baselines described in VI-A. As shown in Fig. 10, our *Sim-Suction-Pointnet (Mask)* achieves the highest success rate across all cluttered categories, with a reliability of **96.67%** for cluttered level 1 objects, **95.00%** for cluttered level 2 objects, and **95.90%** for cluttered mixed objects. *Sim-Suction-Pointnet (Mask)* outperforms *Sim-Suction-Pointnet (No Mask)*, indicating the importance of segmentation masks. These masks provide additional information about instance boundaries, which can be vital in identifying suitable grasping points on the object's surface. When the model has access to this information, it can better focus on the target object and avoid interference from surrounding objects or clutter. The use of segmentation masks also prevents the policy from repeatedly attempting the same unsuccessful suction pose. Instead, it allows the policy to shift its focus to other objects. *Sim-Suction-Pointnet* performs better than *Sim-Suction-Pixelnet*. learning point-wise features from point clouds, which allows the model to focus on local geometric properties and relationships between points. *Sim-Suction-Pointnet* processes raw point cloud data, which is less affected by the domain gap between synthetic and real-world data. On the other hand, *Sim-Suction-Pixelnet* relies on RGB-D images, which are more sensitive to variations in lighting, textures, and other factors that may differ between simulated and real environments. The robot with the Baseline policy takes longer to get prediction results. The executed suction grasp poses from the Baseline policy have collisions with nearby objects in many cases that cause the failure.

*2) **Experiments (picking up specific objects)**:* To evaluate the effectiveness of our method in executing tasks that require

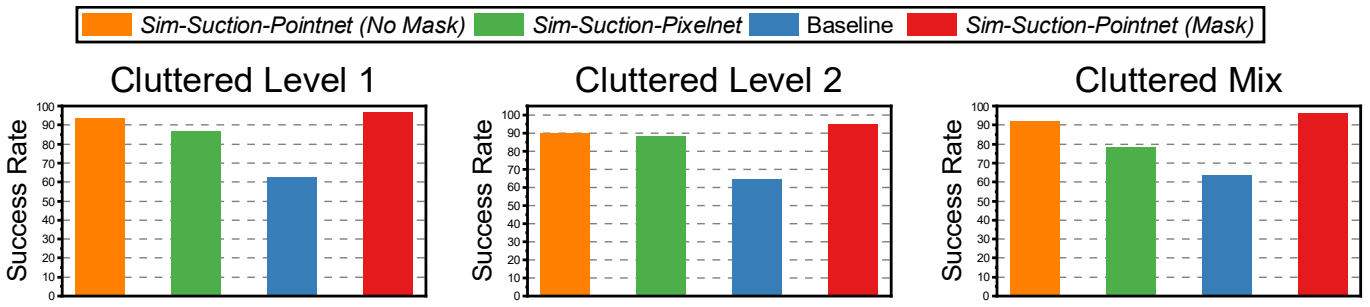


Fig. 10. This figure presents the ablation study results, showing the success rate of attempted grasps in cluttered environments using different methods. **Cluttered Level 1 objects**: *Sim-Suction-Pointnet (No Mask)* achieved a success rate of 93.33% (56 successes in 60 attempts), *Sim-Suction-Pixelnet* 86.89% (53 successes in 61 attempts), the Baseline method 62.50% (35 successes in 56 attempts), and *Sim-Suction-Pointnet (Mask)* 96.67% (58 successes in 60 attempts). **Cluttered Level 2 objects**: *Sim-Suction-Pointnet (No Mask)* achieved a success rate of 90.00% (36 successes in 40 attempts), *Sim-Suction-Pixelnet* 88.57% (31 successes in 35 attempts), the Baseline method 64.52% (20 successes in 31 attempts), and *Sim-Suction-Pointnet (Mask)* 95.00% (38 successes in 40 attempts). **Cluttered Mixed Level 1 and Level 2 objects**: *Sim-Suction-Pointnet (No Mask)* achieved a success rate of 91.84% (45 successes in 49 attempts), *Sim-Suction-Pixelnet* 78.26% (36 successes in 46 attempts), the Baseline method 63.64% (21 successes in 33 attempts), and *Sim-Suction-Pointnet (Mask)* 95.92% (47 successes in 49 attempts).

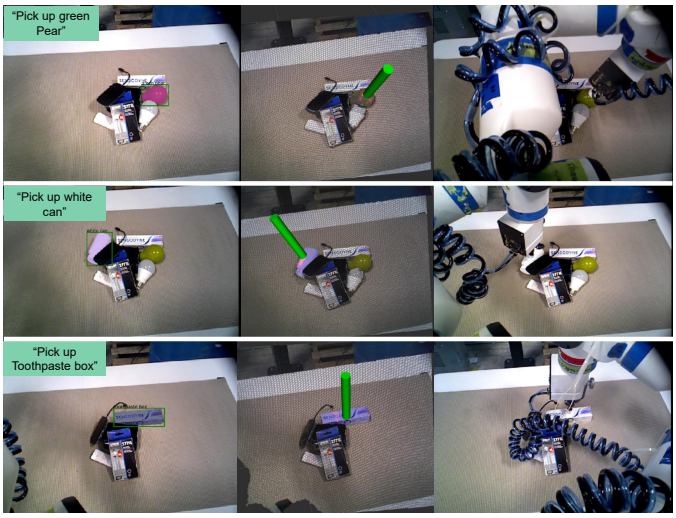


Fig. 11. Qualitative results of experiments for picking up specific objects. The figure displays various instances where the *Sim-Suction-Pointnet* policy successfully identifies and grasps the target object in cluttered environments.

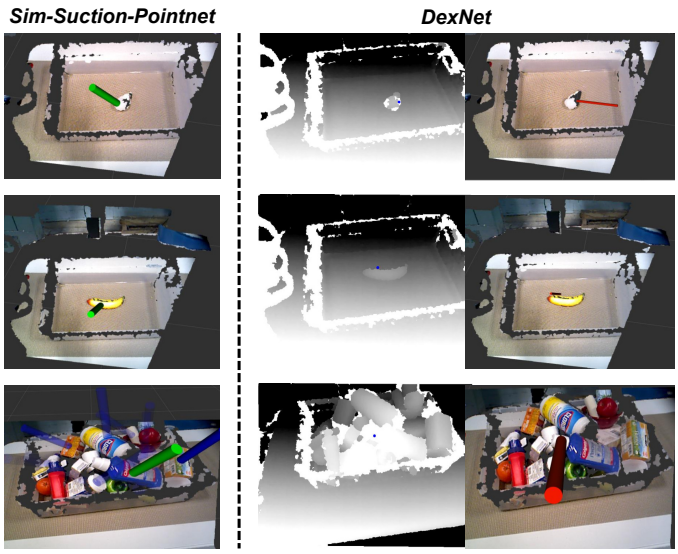


Fig. 12. Qualitative comparison of *Sim-Suction-Pointnet* **(Left)** with DexNet **(Right)**. Top-Row: DexNet generates suction poses on the object edge. Middle-Row: DexNet generates suction poses on the nearby ground. Bottom-Row: DexNet generates suction poses on the unsuctionable area of the object.

selecting specific novel objects (Fig.11), we create cluttered environments using a small set of 8 objects. In these scenarios, the object of interest may be either visible to the camera or hidden beneath a pile of other items. The policy's objective is to search for and locate the target object, successfully grasp it, and complete the task. As the method is sensitive to the text prompt, we perform pre-tests and refine the text prompt to generate a reasonable description for the novel objects of interest. We conduct 20 experiments to pick up specific objects, and the policy achieves a success rate of 16/20. The failure cases occur when the robot picks up other objects due to false detection of the novel objects, as these objects are very similar and lack textures. Further discussion on this issue can be found in the Grounding DINO paper [38].

*3) **Comprehensive Experiments (picking up all objects)**:* To better evaluate and quantify the reliability of *Sim-Suction* in cluttered environments, we perform around 600 attempts and compare them with the state-of-the-art DexNet. By comparing our method with the state-of-the-art, we can establish a benchmark for future research in this area. This comparison allows us to measure the progress made by our method and identify areas where further improvements can be made. Table IV showcases the experimental outcomes for successful attempts against total attempts in cluttered environments for various methods. DexNet-4.0 (GQ-CNN) is trained on synthetic depth images captured from a fixed top-down camera view. Both DexNet and our method aim to test on novel objects that are not present in our respective datasets, emphasizing the zero-shot generalization capabilities. In comparison, our experiments employ a mobile robot with a changing camera viewpoint, which considerably diverges from DexNet's training environment. The *Sim-Suction-Pointnet* (Mask) approach demonstrates superior performance across all cluttered environments, with reliability rates of **96.76%**, **94.23%**, and **92.39%** for cluttered level 1, cluttered level 2, and cluttered mixed objects, respectively. The state-

of-the-art DexNet 4.0-Suction exhibits a lower reliability of 81.22%, 77.73%, and 71.61% for the same scenarios. The results directly reported by DexNet 4.0-Suction [2] are 93%, 80%, and 78%. In DexNet's experiment setup, they chose 50 objects for a cluttered mix environment, while our setup consists of around 60 objects. DexNet performs worse in our experiment setting. One possible reason for this discrepancy is that the experiments conducted in DexNet 4.0 use an industry-level over-bin depth camera, whereas our experiment employs a changing camera view. DexNet 4.0 is also trained on a fixed vision dataset, resulting in suboptimal performance on mobile manipulation platforms. As mentioned by the authors in a seminar [53], Dex-Net 4.0 faces challenges in mobile manipulation platforms with a moving camera that is not mounted on top of the workspace. Figure 12 illustrates several instances where DexNet encounters difficulties. These challenges arise due to noisy depth images, leading DexNet to fail in generating reachable 6D suction poses, which are typically located on the object boundary. Additionally, DexNet employs a segmentation method that only separates the foreground of the scene, rather than employing instance segmentation. As a result, it struggles to handle individual objects when they are placed in cluttered environments. The results emphasize the effectiveness and adaptability of the *Sim-Suction-Pointnet* (Mask) method to various camera perspectives and real-world conditions in intricate cluttered environments. In contrast to the depth images used by DexNet-4.0 (GQ-CNN), *Sim-Suction-Pointnet* (Mask) employs a point cloud-based strategy and uses synergy with a zero-shot RGB segmentation method. This enables *Sim-Suction-Pointnet* (Mask) to be more resilient and adaptable to different camera angles and novel objects. By using the segmentation mask, *Sim-Suction-Pointnet* (Mask) refines the point cloud input and separates the object of interest from the surrounding clutter. This focus on the target object enhances the model's ability to pinpoint appropriate grasping points. Furthermore, the extensive synthetic *Sim-Suction-Dataset* utilized for training *Sim-Suction-Pointnet* (Mask) encompasses a wide variety of object shapes, sizes, and geometries, as well as provides more accurate ground truth labeling. This diverse dataset contributes to the policy's superior generalization abilities in comparison to DexNet-4.0 (GQ-CNN). Examples of failure cases encountered by *Sim-Suction-Pointnet* during experiments (Fig. 13) are primarily situations where the object is obscured from the camera's view since we use a moving camera instead of a high-resolution fixed vision system above the bin. Other cases arise from the nature of cluttered environments, where the object is not stable and may move, rotate, or roll during the grasping process. Only a few cases are caused by the unsuccessful attempts to grasp an object beneath a pile.

TABLE IV
EXPERIMENTAL RESULTS OF SUCCESS ATTEMPTS VERSUS TOTAL ATTEMPTS IN CLUTTERED ENVIRONMENTS FOR DIFFERENT METHODS

Policy	# Attempts	# Total Attempt Fail.	Success Rate	Objects Grasped/Total
Cluttered level 1 (20 Objects per Scene)				
DexNet-4.0 (GQ-CNN)	213	40	81.22%	173/179
***Sim-Suction-Pointnet* (Mask)**	**185**	**6**	**96.76%**	**179/179**
Cluttered level 2 (40 Objects per Scene)				
DexNet-4.0 (GQ-CNN)	238	53	77.73%	185/197
***Sim-Suction-Pointnet* (Mask)**	**208**	**12**	**94.23%**	**196/197**
Cluttered Mix (60 Objects per Scene)				
DexNet-4.0 (GQ-CNN)	229	65	71.61%	164/172
***Sim-Suction-Pointnet* (Mask)**	**184**	**14**	**92.39%**	**170/172**

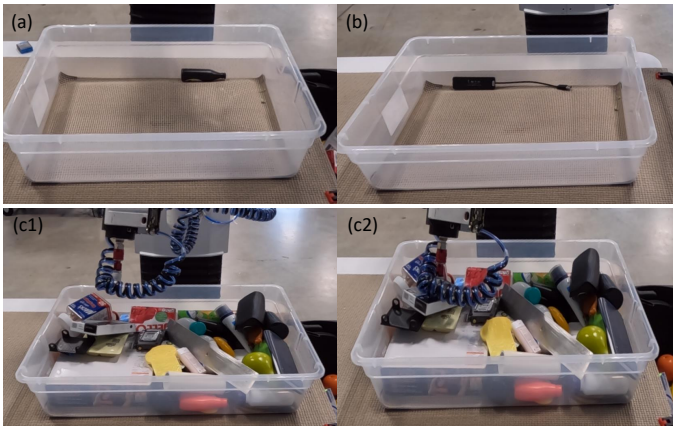


Fig. 13. Examples of failure cases. (a) and (b) The object overlaps with the bin edges. (c1) and (c2) The object is unstable, causing it to move when the robot attempts to form a seal.

VII. CONCLUSIONS & FUTURE WORK

In this paper, we present *Sim-Suction*, a deep learning-based object-aware suction grasp policy for objects in cluttered environments. Experiments conducted on a mobile manipulation platform demonstrate that *Sim-Suction*, learned from the synthetic point cloud dataset *Sim-Suction-Dataset*, achieves a robust success rate in real-world cluttered environments with dynamic viewpoints. It outperforms the state-of-the-art DexNet methods by approximately 21% for mixed cluttered scenes. In the future, we plan to study a multi-gripper grasping policy that enables swapping between different task-specific end-effectors to increase the grasp success rate and handle more challenging objects.

VIII. ACKNOWLEDGEMENTS

The authors would like to acknowledge the use of the facilities at the Indiana Next Generation Manufacturing Competitiveness Center (IN-MaC) for this paper. A portion of this work was supported by a Space Technology Research Institutes grant (# 80NSSC19K1076) from NASA's Space Technology Research Grants Program.

REFERENCES

[1] M. Fujita, Y. Domae, A. Noda, G. A. G. Ricardez, T. Nagatani, A. Zeng, S. Song, A. Rodriguez, A. Causo, I. M. Chen, and T. Ogasawara, "What are the important technologies for bin picking? technology analysis of robots in competitions based on a set of performance metrics," *Advanced Robotics*, vol. 34, no. 7-8, pp. 560–574, 2020. [Online]. Available: https://doi.org/10.1080/01691864.2019.1698463

[2] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.aau4984

[3] J. Li, C. Teeple, R. J. Wood, and D. J. Cappelleri, "Modular end-effector system for autonomous robotic maintenance & repair," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 4510–4516.

[4] J. Mahler, M. Matl, X. Liu, A. Li, D. V. Gealy, and K. Goldberg, "Dex-net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8, 2018.

[5] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. C. Dafle, R. Holladay, I. Morena, P. Qu Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, and A. Rodriguez, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 3750–3757.

[6] H. Cao, H.-S. Fang, W. Liu, and C. Lu, "Suctionnet-1billion: A large-scale benchmark for suction grasping," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8718–8725, 2021.

[7] P. Jiang, J. Oaki, Y. Ishihara, and J. Ooga, "Multiple-object grasping using a multiple-suction-cup vacuum gripper in cluttered scenes," 2023.

[8] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, 2008. [Online]. Available: https://doi.org/10.1177/0278364907087172

[9] A. Morales, E. Chinellato, A. Fagg, and A. P. del Pobil, "Using experience for assessing grasp reliability." *I. J. Humanoid Robotics*, vol. 1, pp. 671–691, 12 2004.

[10] S. Bell and K. Bala, "Learning visual similarity for product design with convolutional neural networks," *ACM Transactions on Graphics*, vol. 34, pp. 98:1–98:10, 07 2015.

[11] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1316–1322, 2015.

[12] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," 2015. [Online]. Available: https://arxiv.org/abs/1509.06825

[13] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," *2011 IEEE International Conference on Robotics and Automation*, pp. 3304–3311, 2011.

[14] P. Jiang, J. Oaki, Y. Ishihara, J. Ooga, H. Han, A. Sugahara, S. Tokura, H. Eto, K. Komoda, and A. Ogawa, "Learning suction graspability considering grasp quality and robot reachability for bin-picking," *Frontiers in Neurorobotics*, vol. 16, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2022.806898

[15] H. Zhang, J. Peeters, E. Demeester, and K. Kellens, "A cnn-based grasp planning method for random picking of unknown objects with a vacuum gripper," *Journal of Intelligent & Robotic Systems*, vol. 103, 12 2021.

[16] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang, "Pointnetgpd: Detecting grasp configurations from point sets," *2019 International Conference on Robotics and Automation (ICRA)*, pp. 3629–3635, 2018.

[17] C. Eppner, A. Mousavian, and D. Fox, "A billion ways to grasp: An evaluation of grasp sampling schemes on a dense, physics-based grasp data set," 2019. [Online]. Available: https://arxiv.org/abs/1912.05604

[18] B. A. Whitehead, "James j. gibson: The ecological approach to visual perception. boston: Houghton mifflin, 1979, 332 pp," *Behavioral Science*, vol. 26, no. 3, pp. 308–309, 1981. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/bs.3830260313

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2015, pp. 3431–3440. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298965

[20] Z. Xu, Z. He, and S. Song, "Umpnet: Universal manipulation policy network for articulated objects," 2022.

[21] W. Zhai, H. Luo, J. Zhang, Y. Cao, and D. Tao, "One-shot object affordance detection in the wild," *International Journal of Computer Vision*, vol. 130, pp. 2472 – 2500, 2021.

[22] X. Chen, T. Liu, H. Zhao, G. Zhou, and Y.-Q. Zhang, "Cerberus transformer: Joint semantic, affordance and attribute parsing," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19 617–19 626, 2021.

[23] A. Gouda and M. Roidl, "Dounseen: Zero-shot object detection for robotic grasping," *ArXiv*, vol. abs/2304.02833, 2023.

[24] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 77–85.

[25] C. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *NIPS*, 2017.

[26] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," 2019. [Online]. Available: https://arxiv.org/abs/1905.10520

[27] P. Ni, W. Zhang, X. Zhu, and Q. Cao, "Pointnet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds," 2020. [Online]. Available: https://arxiv.org/abs/2003.09644

[28] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 441–11 450.

[29] K. Mo, Y. Qin, F. Xiang, H. Su, and L. J. Guibas, "O2o-afford: Annotation-free large-scale object-object affordance learning," *ArXiv*, vol. abs/2106.15087, 2021.

[30] B. Wen, W. Lian, K. E. Bekris, and S. Schaal, "Catgrasp: Learning category-level task-relevant grasping in clutter from simulation," *2022 International Conference on Robotics and Automation (ICRA)*, pp. 6401–6408, 2021.

[31] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "Unseen object instance segmentation for robotic environments," *IEEE Transactions on Robotics*, vol. 37, pp. 1343–1359, 2020.

[32] X. Liu, Y. Zhang, and D. Shan, "Unseen object few-shot semantic segmentation for robotic grasping," *IEEE Robotics and Automation Letters*, vol. 8, pp. 320–327, 2023.

[33] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13 452–13 458, 2021.

[34] S. Ainetter, C. Böhm, R. Dhakate, S. Weiss, and F. Fraundorfer, "Depth-aware object segmentation and grasp detection for robotic picking tasks," *ArXiv*, vol. abs/2111.11114, 2021.

[35] Y. Xiang, C. Xie, A. Mousavian, and D. Fox, "Learning rgb-d feature embeddings for unseen object instance segmentation," in *Conference on Robot Learning*, 2020.

[36] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[37] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. B. Girshick, "Segment anything," *ArXiv*, vol. abs/2304.02643, 2023.

[38] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. yue Li, J. Yang, H. Su, J.-J. Zhu, and L. Zhang, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *ArXiv*, vol. abs/2303.05499, 2023.

[39] Q. Shao, J. Hu, W. Wang, Y. Fang, W. Liu, J. Qi, and J. Ma, "Suction grasp region prediction using self-supervised learning for object picking in dense clutter," *2019 IEEE 5th International Conference on Mechatronics System and Robots (ICMSR)*, pp. 7–12, 2019.

[40] M. Han, W. Liu., Z. Pan, T. Xue, Q. Shao, J. Ma, and W. Wang, "Object-agnostic suction grasp affordance detection in dense cluster using self-supervised learning.docx," 2019. [Online]. Available: https://arxiv.org/abs/1906.02995

[41] J. Liang, V. Makoviychuk, A. Handa, N. Chentanez, M. Macklin, and D. Fox, "Gpu-accelerated robotic simulation for distributed reinforcement learning," 2018. [Online]. Available: https://arxiv.org/abs/1810.05762

[42] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository," *CoRR*, vol. abs/1512.03012, 2015. [Online]. Available: http://arxiv.org/abs/1512.03012

[43] B. Çalli, A. Walsman, A. Singh, S. S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: The YCB object and model set and benchmarking protocols," *CoRR*, vol. abs/1502.03143, 2015. [Online]. Available: http://arxiv.org/abs/1502.03143

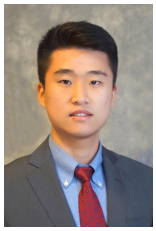
[44] "Omniverse platform for virtual collaboration." [Online]. Available: https://www.nvidia.com/en-us/omniverse/

[45] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg, "Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 1957–1964.

[46] T. Akenine-Möller, E. Haines, N. Hoffman, A. Pesce, M. Iwanicki, and S. Hillaire, *Real-Time Rendering 4th Edition*. Boca Raton, FL, USA: A K Peters/CRC Press, 2018.

[47] O. D. Team, "Openpcdet: An open-source toolbox for 3d object detection from point clouds," https://github.com/open-mmlab/OpenPCDet, 2020.

[48] E. Olson, "Apriltag: A robust and flexible visual fiducial system," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 3400–3407.
[49] C. Eppner, A. Mousavian, and D. Fox, "A billion ways to grasp: An evaluation of grasp sampling schemes on a dense, physics-based grasp data set," in *Robotics Research*, T. Asfour, E. Yoshida, J. Park, H. Christensen, and O. Khatib, Eds. Cham: Springer International Publishing, 2022, pp. 890–905.
[50] M. Macklin, K. Storey, M. Lu, P. Terdiman, N. Chentanez, S. Jeschke, and M. Müller, "Small steps in physics simulation," *Proceedings of the 18th annual ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2019.
[51] N. D. Ratliff, J. Issac, D. Kappler, S. Birchfield, and D. Fox, "Riemannian motion policies," 2018. [Online]. Available: https://arxiv.org/abs/1801.02854
[52] S. Xie and Z. Tu, "Holistically-nested edge detection," *International Journal of Computer Vision*, vol. 125, pp. 3–18, 2015.
[53] K. Goldberg, "The new wave in robot grasping," 2019. [Online]. Available: https://www.youtube.com/watch?v=ATDrSWZXuwk



Juncheng Li received his B.E. degree in Mechanical Engineering from Stony Brook University, NY in 2018. He then obtained his M.S.E. degree in Robotics from the GRASP Lab at the University of Pennsylvania, Philadelphia, PA in 2020. Currently, he is pursuing his Ph.D. in Mechanical Engineering at the Multi-Scale Robotics and Automation Lab (MSRAL) at Purdue University, West Lafayette, IN. His research interests span grasping policy, modular end-effector systems, and computer vision.

David J. Cappelleri (M'09) received a B.S. degree in mechanical engineering from Villanova University, Villanova, PA, USA, an M.S. degree in mechanical engineering from Pennsylvania State University, State College, PA, USA, and a Ph.D. degree in mechanical engineering and applied mechanics from the University of Pennsylvania, Philadelphia, PA, USA. He is currently a Professor with the School of Mechanical Engineering and Weldon School of Biomedical Engineering (By Courtesy) at Purdue University, West Lafayette, IN, USA and directs the Multi-Scale Robotics and Automation Lab.

Prof. Cappelleri is a recipient of the National Science Foundation CAREER Award, the Harvey N. Davis Distinguished Assistant Professor Teaching Award, the Association for Lab Automation Young Scientist Award, and the B.F.S. Schaefer Scholar Award. He is a member of the IEEE Robotics and Automation Society Technical Committee on Micro/Nano Robotics and Automation and is on the Editorial Board of IEEE Robotics and Automation Letters, and the Journal of Micro and Bio Robotics.