SGSIN: Simultaneous Grasp and Suction Inference Network via Attention-Based Affordance Learning

Wenshuo Wang, *Student Member, IEEE*, Haiyue Zhu, *Member, IEEE*, and Marcelo H. Ang Jr, *Senior Member, IEEE*

***Abstract*—Universal handling of a wide and diverse range of objects is a grand and long-lasting challenge for robot manipulation. While recent methods have achieved competitive results in grasping a specific set of objects, it still performs unsatisfactorily when faced with cluttered and diverse objects with different characteristics such as material, shape, texture, etc. One critical bottleneck stems from the limited applicability of single gripper modality, which is not capable of handling the complex real-world tasks. In this paper, we propose a unified grasping inference framework for multi-modality grasping, i.e., the Simultaneous Grasp and Suction Inference Network (SGSIN). SGSIN utilizes the 3D scene input to simultaneously predict feasible grasp poses for multiple modalities of grippers as well as to determine the optimal primitive based on a gripper selection module. SGSIN leverages a novel point-level affordance metric to indicate the potential probability of grasping success for each gripper in a unified manner. A novel backbone network is developed to extract strong 3D feature representations, where the residual block with point and channel attentions is proposed. Our multi-modality framework is trained and evaluated on the GraspNet-1B dataset, and the performances achieve the state of the art on respective grasp and suction tasks. Furthermore, we also evaluate the developed pipeline in real-world tasks with cluttered environments. The state-of-the-art performance demonstrates the effectiveness of our proposed framework.**

***Index Terms*—Affordance learning, attention block, grasp inference, gripper, suction.**

I. Introduction

Autonomous robotic grasping remains a significant challenge due to the complexities involved in adapting to diverse environments. Benefiting from the rapid development of deep learning and computer vision techniques in recent years, research on learning-based grasping [1]–[3] aims to directly predict the grasp strategies using deep neural networks. Previous analytic approaches can be integrated to generate ground-truth labels for network supervision. With the ground truth obtained by human annotation [4] or self-supervision process [5], learning-based methods have the advantage of not requiring prior knowledge of objects, and therefore, have the ability to generalize to unknown or unseen objects. The growing data availability, better computational resources, and more effective algorithms bestow the learning-based methods with significant potentials in robotic grasping.

In recent years, vision-based grasping research has focused on using either 2D RGB(D) images or 3D scene point clouds as inputs to predict feasible grasp poses. The literature [6] offers comprehensive insights and guidelines for benchmarking robotic grasping systems. Previous works can be roughly classified into two categories, i.e., 2D planar and 6-DoF grasp estimations [7]. The former is constrained to grasp from a top-down direction while the latter has no restrictions on the degrees of freedom. 2D planar grasp methods are formulated as the task of evaluating grasp contact points [8], [9] or oriented rectangles [10], [11], which usually follow the pipeline that first samples candidate grasps in formats of either pixels or rectangle anchors, and use analytical or deep learning-based methods to assess the success possibility of a sampled grasp. Gradually, increasing efforts have shifted towards 6-DoF grasping as it allows robots to grasp objects with fewer constraints. In the work [12], the authors consider grasp samples and learn a function to estimate the quality of a grasp candidate. Other approaches [13] directly estimates a single 6-DoF grasp pose and the work [14] reduce the degrees of freedom initially due to the difficulty of generalizing to high-dimensional space. Literature [15] first generates orthographic views of the given objects, which are then used to estimate pixel-wise grasp synthesis for each object. Built upon the principle of the vision-based grasping research, our method analyzes the captured images to make decisions about how to grasp objects effectively.

In order to grasp a wide range of objects in a real-world complex environment, multiple modalities of grippers are beneficial or even indispensable in practice. Parallel-jaw grippers and suction cup grippers are two common types of end-effectors used in robotics for grasping and manipulation. The parallel-jaw grippers consist of two opposing jaws that can be adjusted to accommodate a wide range of objects by

Manuscript received 3 March 2024; revised 19 April 2024 and 19 June 2024; accepted 11 August 2024.

This research is supported by A*STAR "RIE2025 IAF-PP Advanced ROS2-native Platform Technologies for Cross sectorial Robotics Adoption (M21K1a0104)" programme.

Wenshuo Wang is with the Department of Mechanical Engineering, National University of Singapore, 117575, Singapore, (e-mail: wenshuo_wang@u.nus.edu).

Haiyue Zhu is with the Singapore Institute of Manufacturing Technology, Agency for Science, Technology and Research (A*STAR), 138634, Singapore, (e-mail: zhu_haiyue@simtech.a-star.edu.sg).

Marcelo H. Ang Jr is with Advanced Robotics Centre at National University of Singapore, Singapore 117608, Singapore, (e-mail: mpeangh@nus.edu.sg).

clamping the jaws around the object. The suction cup grippers typically use a vacuum system to attach to the surface of an object. It creates a vacuum seal and suction force between the cup and the object's surface. Due to the different mechanisms, parallel-jaw grippers offer versatility and adaptability for various objects, while suction cup grippers are suitable for gripping flat and smooth surfaces with vacuum-based mechanisms. Relying on a single modality may have a negative impact on grasping stability when dealing with various objects. For instance, the widely-used parallel-jaw gripper struggles with grasping objects that have flat surfaces while the suction gripper can easily accomplish this task [16]. Additionally, soft and dexterous grippers are highly adaptive to wrapping and grasping objects with irregular shapes. Therefore, the design of grasping strategy has to consider the combination of multiple grippers, allowing individual modalities to cooperate together to effectively handle grasping in a complex environment.

However, most existing works primarily focus on single-modality grasping, and only a few studies have been conducted on grasping with multiple modalities. Literature [17] presents a novel grasp planning algorithm for hybrid grippers (two-finger grasps, single or double suction grasps, and magnetic grasps). But it requires 6D pose estimation or material segmentation network to satisfy the grasp inference. Literature [8] uses Fully Convolutional Network (FCN) to learn dense pixel-wise affordance map for four pre-defined motion primitives (two for parallel-gripper and two for suction cup gripper). A task planner then executes the primitive with the highest affordance. However, the framework [8] requires experienced users to manually label pixels on RGB-D images, which can be tedious and time-consuming. In another approach, separate GQ-CNNs [18] is trained for each gripper on the Dex-Net 4.0 dataset. Actions are planned by maximizing the predicted grasp quality. Yet, GQ-CNNs relies on providing initial grasp candidates for refinement, which may result in suboptimal final results. In addition, both of the aforementioned works utilize 2D inputs, i.e., RGB-D or depth images, potentially ignoring local geometry information when compared to the representation of 3D data, such as point cloud or voxels.

In this work, we present an integrated grasping inference framework that supports multi-modality grippers with the goal of achieving universal grasping in a complex environment. Our framework addresses two main issues, i.e., 1) identifying the respective grasp poses for each gripper, and 2) determining the optimal primitive. To tackle these challenges, we propose a unified inference framework that can detect feasible grasp poses in parallel and a gripper selection policy to recommend the suitable gripper for each action step. In particular, two most-common types of grippers are considered in this work, i.e., parallel-jaw gripper and suction cup gripper.

When it comes to grasp pose estimation, different end-effectors typically have inconsistent models, making it challenging to unify the pipelines for predicting grasping poses across different grippers. Therefore, we formulate the multi-modality task as a unified two-stage problem, i.e., *where to grasp* — determine grasp locations and *how to grasp* — identify grasp orientations. For the first stage, we propose a novel point-level affordance metric to score each point indicating the probability of grasping success. Two kinds of affordances are defined respectively to describe the graspable and suctionable probability of each point in the scene point cloud. The points with the higher potential probability will be shortlisted in the training and inference process. Furthermore, a novel grasp detection backbone network is developed to extract both global and local features with strong representation ability. The residual block with self-attention mechanisms is integrated to achieve the multi-task inference.

In summary, our contributions are summarized as follows:

- A framework is developed to determine the suitable gripper for each object and identify the grasp poses.
- A novel point-level affordance is proposed to score each point indicating potential probability of grasping success for each type of gripper.
- A novel backbone network with point and channel attention is designed for enhanced point cloud learning.
- Highly competitive results are achieved in both the GraspNet-1Billion benchmark and real world settings.

II. Methodology

A. Problem Formulation

Towards the universal grasping, the robotic pick-and-place problem is considered. Single-modality gripper may not be capable of handling diverse objects due to their different physical properties (i.e. mass, shape or material). Therefore, in this work, an universal grasping pipeline is to be developed to 1) optimally select one suitable gripper as well as 2) predict the corresponding grasp configurations. Without loss of generality, as the parallel-jaw grippers and the suction-based grippers are the most-widely used handlers, our solution focuses on these two grippers in this work.

Specifically, assume that a robot is available to determine a grasping strategy from the action set $\{(\boldsymbol{C}_{\mathbf{g}}|\mathbf{g}), (\boldsymbol{C}_{\mathbf{s}}|\mathbf{s})\}$, where $\mathbf{g}$ and $\mathbf{s}$ denote the selection of parallel-jaw gripper and suction gripper; $\boldsymbol{C}_{\mathbf{g}} = (\boldsymbol{T}_{\mathbf{g}}, \boldsymbol{R}_{\mathbf{g}}, w_{\mathbf{g}})$ and $\boldsymbol{C}_{\mathbf{s}} = (\boldsymbol{T}_{\mathbf{s}}, \boldsymbol{r}_{\mathbf{s}})$ are the grasp and suction configurations, respectively. Here, $\boldsymbol{T}_{\mathbf{g}} \in \mathbb{R}^3$ is the position of grasp center, $\boldsymbol{R}_{\mathbf{g}} \in \mathbb{R}^{3\times3}$ is the orientation matrix of grasp pose, $w_{\mathbf{g}} \in \mathbb{R}$ denotes the the opening degree of the gripper, $\boldsymbol{T}_{\mathbf{s}} \in \mathbb{R}^3$ is the suction center position, and $\boldsymbol{r}_{\mathbf{s}} \in \mathbb{R}^3$ denotes the orientation vector of suction pose. Following GraspNet-1Billion [19], the grasp orientation matrix $\boldsymbol{R}_{\mathbf{g}}$ is decoupled into three subtasks including classifications of viewpoint $\widehat{V}$, in-plane rotation $\widehat{A}$ and approaching distance $\widehat{D}$. In this work, the multi-modality grasping inference problem is formulated as follows, given one point cloud $\boldsymbol{P} \in \mathbb{R}^{N\times3}$ representing the scene observation, the robot is aimed to select the most suitable gripper $u \in (\mathbf{g}, \mathbf{s})$ for next action as well as predict its specific action configuration $(\boldsymbol{C}_u|u)$, which guarantees the success of collision-free picking-up operation.

B. System Overview

Fig. 1 illustrates the pipeline of our proposed framework. In this work, the multi-modality grasping inference pipeline is divided into four modules, i.e., Backbone Network, Affordance Learning, Poses Generation, and Gripper Selection Policy.

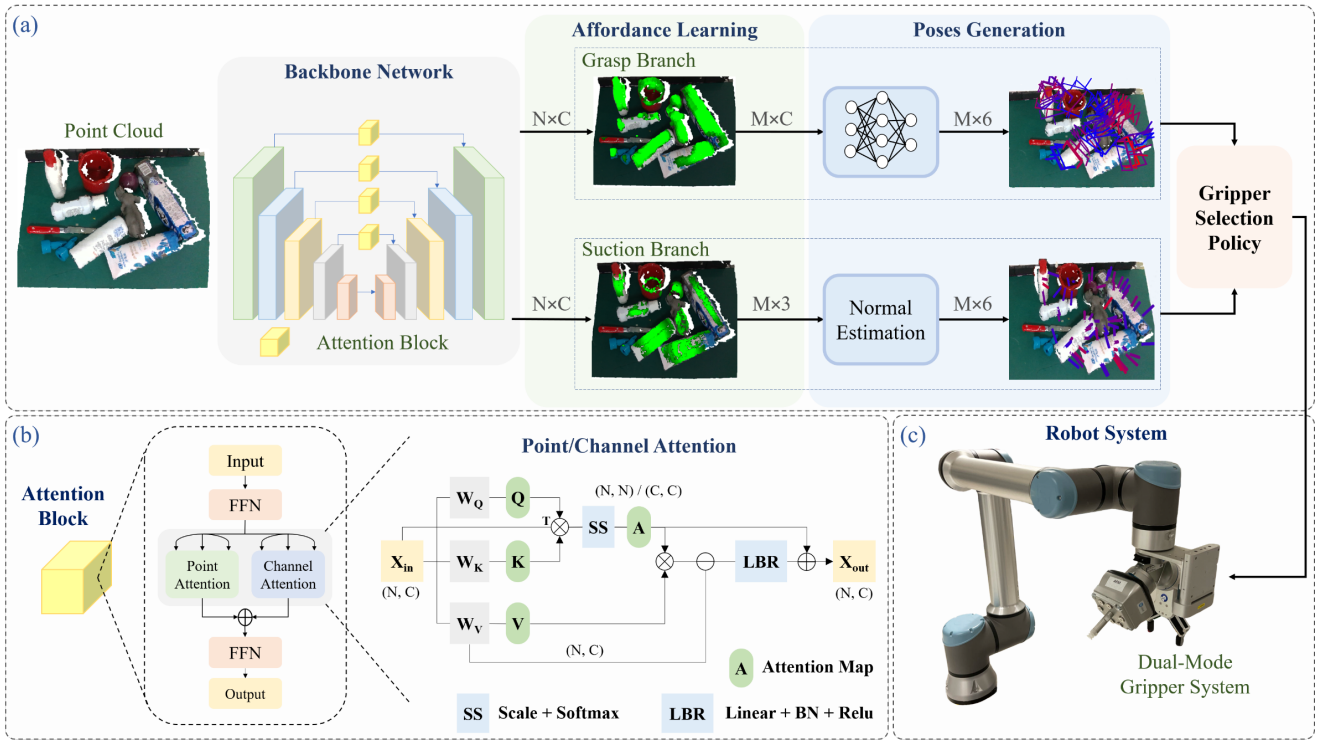


Fig. 1. Overview of SGSIN pipeline. (a) The proposed backbone network first extracts point cloud representations of a scene observation. The Affordance Learning module utilizes the extracted features to output affordances for both grasp and suction tasks. The Pose Generation module then predicts the respective grasp and suction configurations. Based on the final gripper selection policy, the robot will executate the most suitable configuration. (b) The attention block is illustrated, comprising the Point Attention Block and the Channel Attention Block. (c) The real robotic system features a dual-mode gripper system.

Firstly, the Backbone Network $\mathcal{B}(\cdot)$ operates on the scene point cloud $\boldsymbol{P} \in \mathbb{R}^{N\times 3}$ to extract the deep features, denoted as,

$$\boldsymbol{F} = \mathcal{B}(\boldsymbol{P}), \tag{1}$$

where $\boldsymbol{F} \in \mathbb{R}^{N\times C}$ is the point-wise feature. Next, the Affordance Learning module $\mathcal{A}(\cdot)$ addresses the task *where to grasp*, which takes the extracted feature $\boldsymbol{F}$ to output the affordances for both grasp and suction, i.e.,

$$\boldsymbol{A}_{\mathbf{g/s}} = \mathcal{A}(\boldsymbol{F}), \tag{2}$$

where $\boldsymbol{A}_{\mathbf{g/s}}$ represents both grasp affordance $\boldsymbol{A}_{\mathbf{g}} \in \mathbb{R}^{N\times 1}$ and suction affordance $\boldsymbol{A}_{\mathbf{s}} \in \mathbb{R}^{N\times 1}$, which is analogous to the heat map in point cloud. The higher value in $\boldsymbol{A}_{\mathbf{g/s}}$ indicates that this point is more suitable to its respective mode $\mathbf{g/s}$, either grasp or suction. The seed point proposals $\widetilde{\boldsymbol{P}}_{\mathbf{g/s}}$ for further pose inference are shortlisted by applying a threshold on $\boldsymbol{A}_{\mathbf{g/s}}$, followed by the Farthest Point Sampling (FPS) operation.

$$\widetilde{\boldsymbol{P}}_{\mathbf{g/s}} = FPS\big(\boldsymbol{PS}[\boldsymbol{A}_{\mathbf{g/s}} > T_{\mathbf{g/s}}]\big), \tag{3}$$

where $T_{\mathbf{g/s}}$ is the respective threshold for $\mathbf{g/s}$. $\boldsymbol{PS}[\cdot]$ represent a notation used for point selection, where the condition inside the brackets $[\cdot]$ specifies the criteria for selecting points. After that, the Pose Generation module $\mathcal{PG}(\cdot)$ predicts the respective grasp and suction configurations for those shortlisted seed point proposal $\widetilde{\boldsymbol{P}}_{\mathbf{g/s}}$, i.e.,

$$\boldsymbol{C}_{\mathbf{g/s}} = \mathcal{PG}\big(\boldsymbol{F}[\widetilde{\boldsymbol{P}}_{\mathbf{g/s}}]\big). \tag{4}$$

Finally, given the grasp and suction configurations on seed proposals, an Gripper Selection Policy module is employed to determine the most suitable gripper $u^* \in (\mathbf{g}, \mathbf{s})$ for current scene $\boldsymbol{P}$, denoted as

$$u^* = \underset{u\in(\mathbf{g},\mathbf{s})}{\operatorname{argmax}} \left\{ \max_{p\in\widetilde{\boldsymbol{P}}_u} \boldsymbol{A}_u^p \right\}. \tag{5}$$

where $\boldsymbol{A}_u^p$ represents the corresponding affordance of shortlisted seed point proposals. The final predicted configuration $(\boldsymbol{C}_{u^*}|u^*)$ will be executed by robot as the best next action.

C. Backbone Network

In order to learn effective point-wise features for the success of downstream task-oriented modules, we propose a novel full-resolution backbone network $\mathcal{B}(\cdot)$, which adopts the ResUnet architecture built upon Minkowski Engine. ResUNet is a fully convolutional neural network that takes the advantage of both the U-Net architecture and the deep residual learning to get high performance with fewer parameters. Inspired by attention mechanism [20], Point Attention Block and Channel Attention Block are developed and incorporated into the residual connections of the backbone. The purpose of attention blocks is to aggregate the long-range spatial and contextual information for the enhancement of feature capacity.

1) Point Attention Block: Point Attention Block models the long-range dependencies over point features according to the spatial correlations. Following the classical self-attention

mechanism proposed in [20], Query $\boldsymbol{Q}$, Key $\boldsymbol{K}$, Value $\boldsymbol{V}$ variables are defined by applying independent transformations $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, $\boldsymbol{W}_v$ to the input feature $\boldsymbol{F}_{in}$, which is denoted as

$$\begin{aligned} &(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{F}_{in} \cdot (\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v) \\ &\boldsymbol{Q}, \boldsymbol{K} \in \mathbb{R}^{N \times d_{q,k}}, \quad \boldsymbol{V} \in \mathbb{R}^{N \times d_v} \\ &\boldsymbol{W}_q, \boldsymbol{W}_k \in \mathbb{R}^{d_v \times d_{q,k}}, \quad \boldsymbol{W}_v \in \mathbb{R}^{d_v \times d_v}, \end{aligned} \tag{6}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$ share the same feature dimension $d_{q,k}$ and $\boldsymbol{V}$ has the dimension d_v. Next, the intermediate self-attention feature map $\boldsymbol{F}_{pa}$ is caculated as

$$\boldsymbol{F}_{pa} = \boldsymbol{M}_p \cdot \boldsymbol{V} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_a}}\right) \cdot \boldsymbol{V}, \tag{7}$$

where $\boldsymbol{F}_{pa} \in \mathbb{R}^{N \times C}$ is calculated by the product of attention weight matrix $\boldsymbol{M}_p \in \mathbb{R}^{N \times N}$ and value $\boldsymbol{V}$. To obtain $\boldsymbol{M}_p$, $\boldsymbol{Q}\boldsymbol{K}^T$ product is first scaled by $1/\sqrt{d_a}$ and then normalized by a $\mathrm{softmax}$ operation. From the above derivation, the Point Attention Block can be formulated as

$$\begin{aligned} \boldsymbol{F}_{out}^{p} &= \mathcal{PA}\left(\boldsymbol{F}_{in}\right) = \mathcal{LBR}\left(\boldsymbol{F}_{pa} - \boldsymbol{V}\right) + \boldsymbol{F}_{in} \\ &= \mathcal{LBR}\left(\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_a}}\right) \cdot \boldsymbol{V} - \boldsymbol{V}\right) + \boldsymbol{F}_{in}, \end{aligned} \tag{8}$$

where $\mathcal{PA}(\cdot)$ refers to the permutation-invariant self-attention operation, $\mathcal{LBR}(\cdot)$ includes $Linear$, $BatchNorm$, and $ReLU$ basic layers. The final output feature map can be regarded as the sum of learned feature transformed from intermediate $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ variables and the initial input feature, which spatially describes and integrates the long-range dependencies.

2) Channel Attention Block: Besides exploring inter-point correlations, it is also valuable to investigate the inter-channel relationships as each channel of a feature map contains abundant yet distinct semantic information. Therefore, we propose the Channel Attention Block to highlight learned features by distributing attention weights across channels. Note that the main difference of Channel Attention Block from the Point Attention Block lies in how attention is applied across different dimensions of the input. Channel-wise attention focuses on capturing relationships between different channels of a feature map by calculating the dot-product of $\boldsymbol{Q}^{\boldsymbol{T}}$ and $\boldsymbol{K}$. Scale and softmax operators are also applied to obtain channel attention map. The formal formulation can be derived as

$$\begin{aligned} \boldsymbol{F}_{out}^{c} &= \mathcal{CA}\left(\boldsymbol{F}_{in}\right) = \mathcal{LBR}\left(\boldsymbol{F}_{ca} - \boldsymbol{V}\right) + \boldsymbol{F}_{in} \\ &= \mathcal{LBR}\left(\mathrm{softmax}\left(\frac{\boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{K}}{\sqrt{d_a}}\right) \cdot \boldsymbol{V} - \boldsymbol{V}\right) + \boldsymbol{F}_{in}, \end{aligned} \tag{9}$$

where $\boldsymbol{F}_{ca}$ represnets the intermediate self-attention feature map. $\mathcal{CA}(\cdot)$ refers to the channel-wise attention operator, which investigates the channel-wise correlations and gives an enhancement of feature representations.

Finally, element-wise sum operation is performed on the outputs to aggregate the long-range spatial and contextual information together. A feed forward network is applied to do a linear transformation as

$$\begin{aligned} \boldsymbol{F}_{out} &= FFN(\boldsymbol{F}_{out}^{p} \oplus \boldsymbol{F}_{out}^{c}) \\ &= FFN(\mathcal{PA}(\boldsymbol{F}_{in}) \oplus \mathcal{CA}(\boldsymbol{F}_{in})), \end{aligned} \tag{10}$$

where symbol $\oplus$ indicates element-wise summation. Noted that the final output of our attention block shares the invariant feature dimension with the input, which allows it to be a plug-and-play module inserted into the residual connection of ResUnet or some other classical backbones.

D. Affordance Learning

In order to addresses the task of *where to grasp*, Grasp Affordance $\boldsymbol{A}_{\mathbf{g}}$ and Suction Affordance $\boldsymbol{A}_{\mathbf{s}}$ are proposed to identify the grasp and suction centers. The concept of affordance indicates the capability and robustness of a seed point as a grasp or suction center within a task-specific primitive. We assume that the graspable or suctionable area can be inferred by the local geometry and contexts of scene point cloud. In the following, we will introduce how $\boldsymbol{A}_{\mathbf{g}}$ and $\boldsymbol{A}_{\mathbf{s}}$ are formulated in details.

1) Grasp Affordance: Grasp Affordance $\boldsymbol{A}_{\mathbf{g}} \in \mathbb{R}^{N \times 1}$ is defined following previous research GSNet [21] to describe the graspable probability of each point in the scene point cloud $\boldsymbol{P}$. In particular, for a point $p_i \in \boldsymbol{P}$, $\boldsymbol{A}_{\mathbf{g}}^{i}$ is expected to be higher if more successful and stable grasps are centered around. Therefore, $\boldsymbol{A}_{\mathbf{g}}^{i}$ can be denoted as the ratio of successful grasps to all grasp candidates $g_k^{i,j}$ by grid sampling along approach vectors $\widehat{V}$, in-plane rotation angles $\widehat{A}$ and gripper depths $\widehat{D}$:

$$\boldsymbol{A}_{\mathbf{g}}^{i} = \frac{\sum_{j=1}^{\widehat{V}} \sum_{k=1}^{\widehat{L}} \mathbf{1}\left(q_k^{i,j} > \bar{q}\right) \cdot \mathbf{1}\left(c_k^{i,j} = 0\right)}{\sum_{j=1}^{\widehat{V}} \sum_{k=1}^{\widehat{L}} \left|g_k^{i,j}\right|}, i = 1, \ldots, N \tag{11}$$

where $\widehat{L}$ is the number of samples along $\widehat{A}$ and $\widehat{D}$, $q_k^{i,j}$ is the grasp quality score measured by force closure metric [22] and the binary label $c_k^{i,j}$ indicates the collision between the grasp $g_k^{i,j}$ and the environment. In addition, a grasp quality threshold $\bar{q}$ is set manually to determine whether a grasp is successful. To allow for easy comparison of values and ensure the consistent training, $\boldsymbol{A}_{\mathbf{g}}$ is further normalized to improve the numerical stability.

2) Suction Affordance: Analogous to $\boldsymbol{A}_{\mathbf{g}}$, suction affordance $\boldsymbol{A}_{\mathbf{s}} \in \mathbb{R}^{N \times 1}$ is defined to determine the suction centers of the scene observation. Following [23], [24], a physical model of the suction cup gripper is utilized to evaluate the seal formation S_{seal} and the wrench resistance S_{wrench}, which measures surface irregularity based on the change of spring length. Additionally, the geometric-based flatness $S_{flatness}$ is included to measure the geometrical properties within the central area of the object surface. The overall suction affordance $\boldsymbol{A}_{\mathbf{s}}$ is defined as the product of these three components, i.e.

$$\boldsymbol{A}_{\mathbf{s}} = S_{seal} \times S_{wrench} \times S_{flat}, \tag{12}$$

As shown in Fig. 2 (Left), we use a quasi-static spring model to simulate the physical interaction between the suction cup gripper and the object surface. There might be a deformation of all springs when the contact seal is formulated. For simplicity, S_{seal} score is only characterized by the change of perimeter springs as

$$S_{seal} = (1 - \Delta l_{max}) \cdot \sigma, \tag{13}$$

where Δl_{max} denotes the maximum rate of the perimeter springs' deformation, σ is a compensating factor of representing a circle as a polygon.

When it comes to the formulation of S_{wrench}, the suction gripper should be able to tolerate certain wrenches caused by the applied force such as gravity or friction. We simplify the wrench score evaluation by only keeping the Elastic Restoring Torque $\tau = (\tau_x, \tau_y)$, which is concerned with the induced torques along the contact x and y axes due to elastic restoring forces in the suction cup material. τ is illustrated in Fig. 2 (Middle) and S_{wrench} can be denoted as

$$S_{wrench} = 1 - \min\left(1, |\tau| / \tau_{max}\right), \tag{14}$$

where τ_{max} is the maximum tolerated torque. From above definitions of S_{seal} and S_{wrench}, the following model-based policy tries to minimize the energy consumption to achieve stable suction by predicted suction poses.

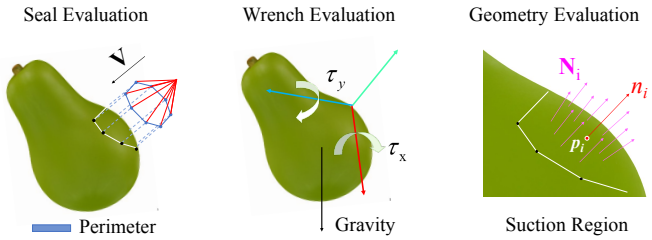


Fig. 2. Components of suction evaluation. (**Left**) the quasi-static spring model with Perimeter springs represented by blue lines. (**Middle**) Wrench evaluation for the suction cup model. (**Right**) Local geometry.

Additionally, S_{flat} is based on the assumption that a suction cup model can achieve a stable grasp when the contact region is flat and smooth. As shown in Fig. 2 (Right), we have a suction center $p_i \in \boldsymbol{P}$ and a normal vector n_i, the k-nearest neighbors algorithm (k-NN) method is used to find K closest neighbors $\boldsymbol{P}_i = \{p_i^k | p_i^k \in \boldsymbol{P}, k = 1, \ldots, K\}$ at p_i and corresponding normal vectors $\boldsymbol{N}_i = \{n_i^k | k = 1, \ldots, K\}$ are also estimated. Then, the curvature of the local suction region can be defined as the average inner product between n_i and $\boldsymbol{N}_i$ by

$$S_{flat}^{i,1} = \frac{1}{K} \sum_{k=1}^{K} \frac{\left| < n_i, n_i^k > \right|}{\|n_i\| \cdot \left\|n_i^k\right\|}. \tag{15}$$

Note that $S_{flat}^{i,1}$ only considers the variance of normal vectors, which may not be sufficient as the robot can not achieve stable suction on the boundary area with a step-like surface. The upper and lower surface may have similar normal vectors but it is intuitively impossible to achieve suction across surfaces due to the existence of the middle gap. Therefore, we also consider the smooth property by evaluating the average perpendicular distance from points in the suction region to their linearly fitted plane. The second term of S_{flat}^{i} can be formulated as

$$S_{flat}^{i,2} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{PD}(p_i^k, \mathrm{Plane}(\cdot)), \tag{16}$$

where $\mathrm{PD}(\cdot)$ denotes the perpendicular distance calculation from a point to a plane. $\mathrm{Plane}(\cdot)$ is the fitted plane equation of suction region. From the above perspective, S_{flat} can be finally formulated as

$$S_{flat}^{i} = \lambda \cdot S_{flat}^{i,1} + (1 - \lambda) \cdot S_{flat}^{i,2}, \tag{17}$$

where λ is a modulating weight factor.

E. Pose Generation

Pose Generation module $\mathcal{PG}(\cdot)$ is proposed to address the task *how to grasp*, grasp and suction pose estimation branches are stacked in parallel to predict grasp orientation matrix $\boldsymbol{R_g}$ and suction orientation vector $\boldsymbol{r_s}$ for each seed point proposal. For the prediction of $\boldsymbol{R_g}$, we follow the common practice [21] to decouple the pipeline into View Selection, Cylinder-Group and Angle-Depth Searching sub-modules. Firstly, the View Selection module $\mathcal{VS}(\cdot)$ operates on the shortlisted seed point proposals $\widetilde{\boldsymbol{P}}_{\mathbf{g/s}}$ to output view-wise probability $\boldsymbol{V'} = \left\{ \boldsymbol{v'_j} \subset [0,1]^{\widehat{V}} \mid j = 1, \ldots, M \right\}$, which can be represented as

$$\boldsymbol{V'} = \mathcal{VS}\big(\boldsymbol{F}[\widetilde{\boldsymbol{P}}_{\mathbf{g/s}}]\big), \tag{18}$$

where $\boldsymbol{F}[\widetilde{\boldsymbol{P}}_{\mathbf{g/s}}]$ denote the respective 3D features of $\widetilde{\boldsymbol{P}}_{\mathbf{g/s}}$. $\boldsymbol{v'_j} \subset [0,1]^{\widehat{V}}$ indicates $\widehat{V}$ dimensional scores of all sampled viewpoints centered on each seed point, based on which one view with highest score is selected as the approach direction. Next, Cylinder Group $\mathcal{CG}(\cdot)$ samples neighbors around M seed points in directional cylinder spaces, which are determined by seed points coordinates and selected view vectors, i.e.,

$$\widetilde{\boldsymbol{P}}_{\mathbf{g}}^{K} = \mathcal{CG}(\widetilde{\boldsymbol{P}}_{\mathbf{g}}, \boldsymbol{V'}), \tag{19}$$

where K denotes the number of samples in each group and $\widetilde{\boldsymbol{P}}_{\mathbf{g}}^{K} \in \mathbb{R}^{M \times K \times 3}$ represents candidates in all groups. Furthermore, all point candidates are transformed in cylinder space to obtain a unified coordinate representation. After that, a combination including a MLP network, a max-pooling layer and a new MLP network is applied on feature candidates $\widetilde{\boldsymbol{F}}_{\mathbf{g}}^{K} \in \mathbb{R}^{M \times K \times C}$ to output seed grasps orientations $\widetilde{\boldsymbol{R}}_{\mathbf{g}}$, i.e.,

$$\widetilde{\boldsymbol{R}}_{\mathbf{g}} = \mathbf{MLPs}(\widetilde{\boldsymbol{F}}_{\mathbf{g}}^{K}), \tag{20}$$

where $\widetilde{\boldsymbol{R}}_{\mathbf{g}} \in \mathbb{R}^{M \times (\widehat{A} \times \widehat{D} \times 2)}$ includes grasp quality scores and widths for all angle-depth combinations. Finally, the candidate with the highest grasp quality is selected as $\boldsymbol{R_g}$, i.e.,

$$\boldsymbol{R_g} = \max(\widetilde{\boldsymbol{R}}_{\mathbf{g}}). \tag{21}$$

For the suction task, the prediction of orientation vector $\boldsymbol{r_s}$ is much simplified and intuitive. The normal of local neighbors at suction center $\boldsymbol{T_s}$ is directly determined as $\boldsymbol{r_s}$.

F. Loss Function

The total loss for our network is composed of L_A and L_P terms which correspond to the Affordance Learning and Pose Generation modules respectively. First of all, L_A can be formulated as a multi-task loss as

$$L_A = \sum_{i=1}^{M} (\lambda_o L_o^i + \sum_{u \in (\mathbf{g}, \mathbf{s})} \lambda_u L_u^i), \tag{22}$$

where L_o^i is a binary cross-entropy loss for the point-level objectness classification, L_u^i denotes the Smooth L1-loss of the regression of affordance values. Compared to grasp affordance

loss $L^i_{\mathbf{g}}$, suction affordance loss $L^i_{\mathbf{s}}$ is composed of three terms based on seal, wrench and flatness evaluation:

$$L^i_{\mathbf{s}} = L^i_{seal} + L^i_{wrench} + L^i_{flat}. \tag{23}$$

From the above perspective, the Pose Generation module predicts the grasp and suction pose estimation in parallel. However, due to the non-learning based method used in suction pose generation branch, L_P only includes loss terms of each subtask for the grasp orientation $\boldsymbol{R}_{\mathbf{g}}$ prediction. Based on Eq. 18, 19, 20 and 21, L_P is computed as

$$L_P = \sum_{i=1}^{M} (\lambda_v L^i_v + \lambda_w L^i_w + \lambda_s L^i_s), \tag{24}$$

where L_v, L_w and L_s represent Smooth L1-loss for the regression of approach vectors, gripper widths and grasp scores respectively. In the training stage, λ_o, λ_u, λ_v, λ_w and λ_s are modulating factors to tune the weights in calculation of L_A and L_P, which formulate the total loss L by the final weighted sum as

$$L = \alpha L_A + \beta L_P. \tag{25}$$

G. Gripper Selection Policy

As illustrated in Eq. 5, a gripper selection policy is defined to determine the optimal primitive in each robotic pick-and-place step. The robot will execute the primitive with the highest affordance value by comparing all point candidates in the scene observation. Although the policy is capable of selecting one primitive, further "calibration" is still required due to the distributional bias between different affordances. The distribution of foreground grasp and suction affordance D_g and D_s of the training set is demonstrated in Fig.3 (Left). It can be observed that the percentage of D_s is generally lower in the the range of intervals. We attribute this range disparity to their varying definitions. In order to balance different affordances ranges so that the robot would not excessively depend on one gripper, a function $\mathcal{M}$ is learned to mitigate the bias so that the transformed distributions $D^{'}_g$ and $D^{'}_s$ can be aligned. Consequently, the definitions of $\mathcal{M}$, $D^{'}_g$ and $D^{'}_s$ are given by

$$\begin{aligned} \mathcal{M} &= \mathcal{QT}_s^{-1} \cdot \mathcal{QT}_g, \\ (D^{'}_g, D^{'}_s) &= (\mathcal{M} \cdot D_g, D_s), \end{aligned} \tag{26}$$

where $\mathcal{QT}_s$ and $\mathcal{QT}_g$ denote the functions learned to map D_g and D_s to the standard Gaussian distribution by applying Quantile Transformer [25]. The function $\mathcal{M}$ first maps D_g to standard Gaussian distribution which is then transformed to D_s by the inverse of $\mathcal{QT}_s$. Fig. 3 (Right) shows that the calibrated grasp affordance shares a similar distribution with the original suction affordance. Given a new scene point cloud in the inference stage, $\mathcal{M}$ is applied to the predicted grasp affordance so that our gripper selection policy becomes more stable when doing the pairwise comparison of affordance values.

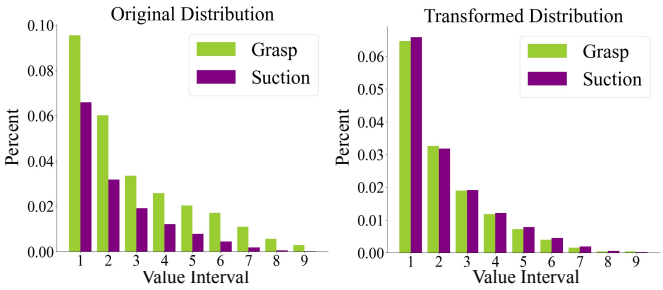


Fig. 3. The value from 0.1 to 1 has been divided into 9 equal intervals. (**Left**) The distributional histograms of affordances before calibration. (**Right**) The distributional histograms of affordances after calibration.

III. Experiments

A. Dataset

In this work, we utilize three datasets: GraspNet-1Billion [19], Suctionnet-1billion [24], and MetaGraspNet [31]. The GraspNet-1Billion dataset contains 88 objects and 190 scenes, with full annotations for object 6D poses, instance masks, camera poses, and 6-DoF grasp poses. All scenes are divided into train, test seen, test similar, and test novel categories, consisting of 100, 30, 30, and 30 scenes, respectively. The dataset includes a total of 97,280 RGB-D images, with 512 images captured for each scene by RealSense and Kinect cameras. The setting of SuctionNet-1Billion is similar to GraspNet-1Billion, but it specifically focuses on suction-based grasping tasks. MetaGraspNet is a large-scale photo-realistic dataset constructed via physics-based metaverse synthesis. Synthetic scenes are generated in metaverse and real world evaluation settings. The MetaGraspNet-Sim and MetaGraspNet-Real datasets provide fully comprehensive ground truth for parallel-jaw and vacuum gripper and scene reasoning. Our experiments with these datasets demonstrated the robustness and generalization capabilities of our method across different object types and grasping conditions.

B. Implementation Details

In our grasp representation, the viewpoint $\widehat{V}$, in-plane rotation $\widehat{A}$ and approaching distance $\widehat{D}$ are sampled to 300, 12 and 4, respectively. As part of our pre-processing techniques, we utilize random downsampling to select a subset of $N = 15k$ points from the raw observation, and outlier removal is employed to filter noisy points. Subsequently, the backbone equipped with an encoder-decoder architecture extracts point-level features with a dimensionality of $C = 512$. In the residual attention branch of the backbone network, $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, $\boldsymbol{W}_v$ are all implemented by a 1-layer MLP. We set the affordance threshold $T_{\mathbf{g/s}}$ as 0.1 and 0.2 to separately shortlist $M = 1024$ seed points. In pose generation module, each cylinder space includes 16 neighbors with radius $r = 0.05$ and height range from -0.02m to 0.04m. Two MLPs are finally used to predict the grasp scores and widths based on grouped proposals. In addition, weights coefficients λ_o, $\lambda_{\mathbf{g}}$, $\lambda_{\mathbf{s}}$, α and β are set to 1, 10, 10, 1 and 1 empirically to compute the overall learning objective. In training stage, we use the Adam optimizer with an initial learning rate of 0.001. The learning

TABLE I
EVALUATION RESULTS FOR DIFFERENT METHODS ON GRASPNET-1B REALSENSE/KINECT.

Metrics	Methods	Seen			Similar			Novel		
		AP	$AP_{0.8}$	$AP_{0.4}$	**AP**	$AP_{0.8}$	$AP_{0.4}$	**AP**	$AP_{0.8}$	$AP_{0.4}$
Grasp Top 1-50	GraspNet-1B [19]	27.56/29.88	33.43/36.19	16.95/19.31	26.11/27.84	34.18/33.19	14.23/16.62	10.55/11.51	11.25/12.92	3.98/3.56
	RGBMatter [26]	27.98/32.08	33.47/39.46	17.75/20.85	27.23/30.40	36.34/37.87	15.60/18.72	12.25/13.08	12.45/13.79	5.62/6.01
	TransGrasp [27]	39.81/35.97	47.54/41.69	36.42/31.86	29.32/29.71	34.80/35.67	25.19/24.19	13.83/11.41	17.11/14.42	7.67/5.84
	DGCAN [28]	48.95/47.32	59.67/57.27	42.24/38.55	41.46/35.73	50.31/44.22	33.69/26.99	17.48/16.10	21.83/20.01	7.90/7.81
	SBG [29]	58.95/–	68.18/–	54.88/–	52.97/–	63.24/–	46.99/–	22.63/–	28.53/–	12.00/–
	HGGD [30]	59.36/60.26	–/–	–/–	51.20/48.59	–/–	–/–	22.17/18.43	–/–	–/–
	GSNet [21]	65.70/61.19	76.25/71.46	61.08/56.04	53.75/47.39	65.04/56.78	45.97/40.43	23.98/19.01	29.93/23.73	14.05/10.60
	Ours	**68.97/63.96**	**81.25/74.82**	**62.25/56.94**	**58.76/51.02**	**72.23/62.36**	**47.84/41.99**	**26.69/20.72**	**33.33/25.89**	**14.53/11.02**
Suction Top 50	Normal STD	10.10/6.16	1.01/0.74	12.64/7.20	19.23/17.54	2.26/2.38	25.95/24.68	3.50/1.34	0.23/0.12	4.22/1.08
	DexNet 3.0 [23]	15.50/9.46	1.53/0.96	20.22/11.90	18.92/14.93	2.62/**2.45**	24.51/19.05	2.62/1.68	0.35/0.04	5.32/1.89
	ARC-Vision [8]	**28.31**/14.82	3.37/1.00	38.58/20.02	26.15/14.23	3.35/1.61	34.83/18.66	7.79/2.85	**0.38**/0.17	9.78/3.35
	SuctionNet [24]	**28.31**/15.94	**3.41/1.25**	38.56/20.89	26.64/18.06	**3.42**/1.96	35.34/24.00	8.23/3.20	0.35/0.23	10.29/3.78
	Ours	28.18/**16.63**	2.53/**1.25**	**38.94/21.15**	**31.48/19.04**	3.36/2.40	**38.72/25.87**	**8.33/3.24**	0.36/**0.26**	**10.41/3.81**

rate is multiplied by 0.95 per epoch. The model is trained on one NVIDIA GTX 3090 GPU with batch size 2. The entire training process takes approximately 40 hours to converge to the optimal configuration. In terms of inference time, it takes 0.1 second to output the best grasp configuration.

C. Results and Evaluation

Fig. 4 illustrates the predicted grasps and suctions by SGSIN specifically for scene 120, 140 and 170 in GraspNet-1Billion dataset. We divide the affordance value range into three groups, resulting in a total of 50 seed proposals chosen as either grasp or suction centers. The color-coded representations of grasps and suctions in the figure correspond to the respective affordance scores of the configurations. The marked colors correspondingly shift from blue to red within the [0,1] range. Compared to solely visualizing top grasps, we are able to effectively depict grasp poses with varying degrees of quality. The figure presented demonstrates reasonable observations of the predicted proposals, thereby validating the robustness and efficacy of our proposed framework.

Table I shows the evaluation results of our SGSIN and several baselines training on GraspNet-1Billion dataset. To measure the predicted grasps on a cluttered scene quantitatively, we adopt the $Precision@k$ metric to measure the precision of top-k ranked grasps. For both grasping and suctioning, the number k is set to 50 and AP_μ indicates the average precision of top-50 ranked grasps. Specifically, AP_μ is determined by the friction coefficient μ ranging from 0.2 to 1.0 with an interval of $\Delta\mu$ = 0.2.

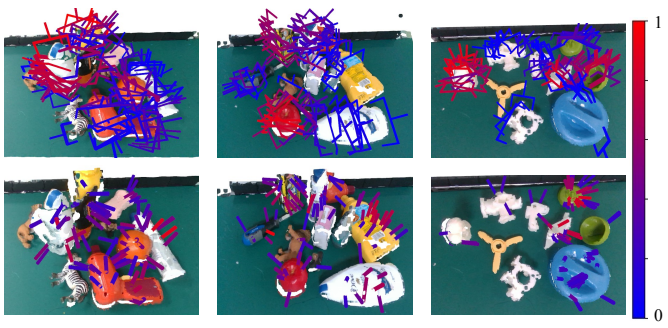


Fig. 4. Visualization examples of predicted grasps (first row) and suctions (second row) performed by our method on GraspNet-1Billion. Grasps are color-coded by the predicted affordance score (higher value in deeper red).

As demonstrated in Table I, notable disparities in terms of metrics are observed between the Realsense and Kinect domains. It can be clearly shown that SGSIN achieves state-of-art performance over all evaluation metrics for the grasping task and most metrics for the suction task. Taking the Realsense metrics as an example, our approach achieves 68.97/63.96%, 58.76/51.02%, and 26.69/20.72% AP for the test seen, test similar, and test novel datasets, respectively, outperforming GSNet which previously exhibited SOTA performance in parallel-jaw gripper grasping. Overall, SGSIN achieves a relative improvement by 3-5 AP on all metrics.

To establish a benchmark for the suction task, we choose SuctionNet [24] as the baseline. Table I also reports the evaluation results using a suction cup gripper. It can be observed that SGSIN outperforms SuctionNet on both cameras of most metrics. For test seen, test similar and test novel categories, SGSIN achieves 28.18/16.63%, 31.48/19.04% and 8.33/3.24% AP for Realsense/Kinect inputs. These improvements can be attributed to the robustness and generalization capabilities of our model, which are derived from the strong representations of 3D point-wise learning. The advancements observed in both grasp and suction tasks confirm the effectiveness of our approach in cluttered robotic grasping environments.

D. Ablation Studies

1) Residual Attention: To investigate the impact of our proposed attention block, we conduct several studies with different backbones for point cloud representation learning. The performances of these backbones in terms of final AP metrics on all test sets are shown in Table II. Compared to PointNet++ [32], ResUnet backbone achieves better performance. It can be observed that our Point Attention Block (PA) and Channel Attention Block (CA) make great contributions,

resulting in performance gains on all test sets including seen, similar and novel. And our attention block with integration of these two blocks further obtain improvements by achieving SOTA on most of AP metrics of grasp and suction tasks. The results illustrated in Table II convincingly validate the effectiveness and benefits of our proposed attention block.

TABLE II
ABLATION STUDY OF DIFFERENT BACKBONES.

Attention Type	AP Seen	AP Similar	AP Novel
PointNet++	60.01 / 22.55	53.83 / 25.85	23.18 / 7.70
ResUnet	66.23 / 27.99	55.68 / 28.54	25.26 / 8.11
ResUnet + PA	67.98 / **28.43**	58.14 / 31.12	25.92 / 8.19
ResUnet + CA	68.21 / 27.43	58.03 / 30.95	26.09 / **8.35**
ResUnet + PCA	**68.97** / 28.18	**58.76** / **31.48**	**26.69** / 8.33

2) Multi-Task Learning: Instead of training separate pipelines to accomplish specific tasks, the multi-modality grasping can be regarded as a typical multi-task learning problem. Table III demonstrates the performance of SGSIN training grasp and suction branch separately and simultaneously. It can be observed that the jointly learning strategy helps achieve an overall improvement of 1-2 AP compared to separately learning strategy. This improvement can be attributed to the shared point-level representations from the same backbone across different branches.

TABLE III
ABLATION STUDY OF MULTI-TASK LEARNING

Training Branch	AP Seen	AP Similar	AP Novel
Grasp/Suction	67.10 / **28.37**	56.97 / 30.16	26.01 / 8.31
Integration	**68.97** / 28.18	**58.76** / **31.48**	**26.69** / **8.33**

3) Test on other dataset: In order to further support the superiority of our method, we have augmented our evaluations based on an additional dataset - MetaGraspNet [31]. We leverage the fundamental architecture of SGSIN and conduct training on the synthetic MetaGraspNet dataset. The whole real MetaGraspNet dataset is employed for testing purposes, where we test all trained models by evaluating their predicted grasps of each scene within the simulation environment built by the open-source Pybullet engine [33]. Both GSNet [21] and SuctionNet [24] trained on the large-scale real-world GraspNet-1Billion are also evaluated against their versions trained on the synthetic MetaGraspNet dataset. We simulate the grasps and use the unified Success Rate (SR) and Completion Rate (CR) to intuitively compare the grasp performance of different models across different datasets. From the results in Table IV, we can see that SGSIN trained on synthetic MetaGraspNet dataset still achieved the best performance of an overall success rate of 78.8%/82.3% and a completion rate of 82.6%/88.5%. These results further validate the effectiveness and advantages of SGSIN.

TABLE IV
EVALUATION RESULTS ON REAL METAGRASPNET

Trained Dataset	Success Rate / Completion Rate (%)		
	GSNet [21]	SNet [24]	SGSIN
GraspNet-1Billion	66.5% / 81.2%	63.8% / 74.7%	**78.8**% / **82.6**%
MetaGraspNet-Syn	71.9% / 84.0%	65.1% / 81.6%	**82.3**% / **88.5**%

∗ SNet is the abbreviation of SuctionNet.

E. Physical Experiments

Besides evaluating on public dataset, we also conducted real-world bin-picking experiments. The objective of physical experiments is to evaluate SGSIN in real-world grasping tasks. To align with the goal of testing the robot's ability to handle different geometries, objects with diverse shapes and sizes are selected. Our selected objects are also commonly found in households or industrial settings. We choose a total of ten objects: box, tape, spoon, dinosaur, vegetable, plate, charm, pipe fittings, poly bag and screwdriver. Then, we create a controlled workspace with appropriate lighting conditions and arrange five objects within the workspace to ensure a variety of placements and orientations to challenge the robot's perception and grasping capabilities. Specifically, we use ROS (Robot Operating System) to provide communication between different components of a robotic system. The operating system is employed at the server computer side with Ubuntu 20.04 LTS and the open-source software framework MoveIt is used for motion planning, kinematics, and control of robotic systems.

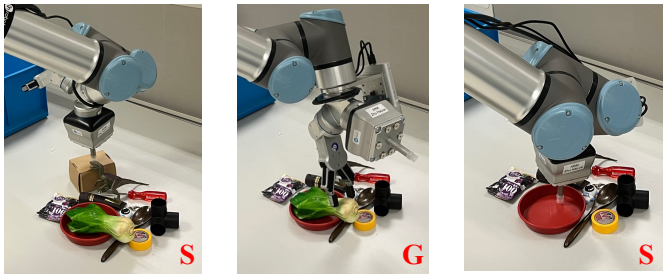


Fig. 5. First three snapshots of the robot grasping objects in a cluttered scene. The red symbols at bottom indicate the primitive the robot execute(G: grasp; S: suction).

Fig. 5 shows the grasping process of a cluttered scene. The robot would select the most suitable gripper based on current observation. In addition, two metrics termed success rate SR and completion rate CR are introduced and defined as the ratio of successful grasps and objects completion. A grasp will be regarded as successful if the object is successfully placed in the bin box, and an object will be considered complete if it is picked successfully within three grasp attempts. Table V reports detailed experimental results for multiple-objects scenes grasping. Compare to GSNet [21] and SuctionNet [24], SGSIN achieved the best performance of an overall success rate of 93.3% and a completion rate of 100%. For most test scenes, it can achieve a nearly 100% SR and CR which further illustrate the robustness and effectiveness of SGSIN.

TABLE V
EXPERIMENTAL RESULTS OF MULTIPLE OBJECTS GRASPING

Object IDs	Success Rate / Completion Rate (%)		
	GSNet [21]	SNet [24]	SGSIN
1, 2, 3, 4, 5	83.3% / 100%	50% / 80%	100% / 100%
1, 4, 5, 6, 8	71.4% / 100%	33.3% / 60%	83.3% / 100%
1, 3, 4, 7, 8	71.4% / 100%	30% / 60%	83.3% / 100%
2, 4, 5, 6, 9	57.1% / 80%	57.1% / 80%	100% / 100%
2, 3, 6, 8, 10	83.3% / 100%	30% / 60%	100% / 100%
All	73.3% / 96%	40.1% / 68%	93.3% / 100%

∗ SNet is the abbreviation of SuctionNet. Objects 1-10: box, tape, spoon, dinosaur, vegetable, plate, charm, pipe fittings, poly bag, screwdriver.

Fig. 6 illustrates the predicted grasps and suctions in real-world scenario. The same color-coded representations, as demonstrated in Section III.C, are applied. As shown in Fig. 6, our grasp model performs relatively well on large-scale and regular objects. However, for small and irregular objects such as pipe fittings, the model performs relatively poorly. This is because smaller or irregular objects typically provide less surface area and fewer potential grasp points, which make it difficult for the model to identify and select optimal grasp configurations. In practice, limitations in real sensor accuracy can also result in false grasps during final execution.

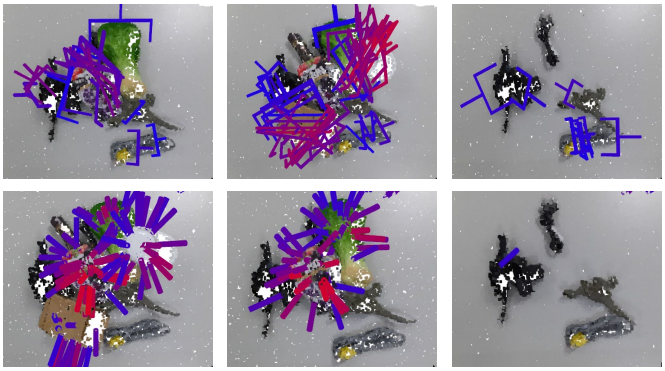


Fig. 6. (**Top Row**) Predicted grasps in real-world scenario. (**Bottom Row**) Predicted suctions in real-world scenario.

In details, our false cases can be divided into false negative grasps and false positive grasps. False negative grasps occur when our robotic system fails to detect a successful grasp. Factors like complex object geometries or inadequate lighting conditions often hinder the accurate detection of graspable points. For instance, when handling small objects like electrical cables, our system struggles to identify successful grasps due to sensor noise. False positive grasps occur when our robotic system erroneously identifies non-graspable points as valid grasps. The Fig. 7 demonstrates some cases the robot fails to execute the predicted grasps. Common scenarios contributing to these failures include inaccurate grasp predictions, collisions between the gripper and the object, and attempts to grasp multiple objects simultaneously. False positive grasps in our robotic system often stem from a combination of factors, including limitations in sensor accuracy, ambiguities in perception caused by variations in object shape, and a lack of contextual or spatial information for guiding grasp planning. To minimize false grasp cases, our aim is to refine our grasp detection algorithm to incorporate advanced feature extraction networks, integrate spatial and semantic information, and leverage machine learning approaches to dynamically adjust grasp thresholds based on real-time feedback.

Fig. 7. False cases of robotic grasping.

IV. CONCLUSION

In this paper, we propose a novel framework named SGSIN to predict 6-DoF grasp poses and determine the optimal primitive for multi-modality grasping. The point-level affordance metric is leveraged to indicate the probability of grasping success and then positive points with high affordance are selected as grasp centers. Furthermore, both point and channel attention mechanisms are incorporated in backbone to enhance 3D scene representations. Exhaustive experiments conducted on Graspnet-1Billion dataset and in real world settings demonstrate the effectiveness of SGSIN for grasping cluttered objects. It's also important to recognize that SGSIN has several limitations, including computational expense and limited compatibility with gripping small or fragile objects. In forthcoming research, we plan to train a more lightweight network that can achieve the same functionality with reduced computational resources. In addition, we aim to research on SGSIN extension with more grasp modalities (e.g. multi-finger gripper or magnetic gripper) that can conform to greater variety of objects.

REFERENCES

[1] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[2] S. Yu, D.-H. Zhai, and Y. Xia, "Egnet: Efficient robotic grasp detection network," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 4, pp. 4058–4067, 2022.

[3] D. Wang, C. Liu, F. Chang, N. Li, and G. Li, "High-performance pixel-level grasp detection based on adaptive grasping and grasp-aware network," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 11, pp. 11 611–11 621, 2021.

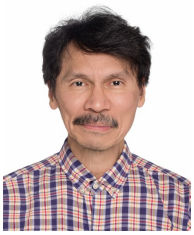
[4] A. Ramisa, G. Alenya, F. Moreno-Noguer, and C. Torras, "Using depth and appearance features for informed robot grasping of highly wrinkled clothes," in *2012 IEEE International Conference on Robotics and Automation*, pp. 1703–1708. IEEE, 2012.

[5] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *2016 IEEE international conference on robotics and automation (ICRA)*, pp. 3406–3413. IEEE, 2016.

[6] S. D'Avella, M. Bianchi, A. M. Sundaram, C. A. Avizzano, M. A. Roa, and P. Tripicchio, "The cluttered environment picking benchmark (cepb) for advanced warehouse automation: Evaluating the perception, planning, control, and grasping of manipulation systems," *IEEE Robotics & Automation Magazine*, 2023.
[7] 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.
[8] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo *et al.*, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," *The International Journal of Robotics Research*, vol. 41, no. 7, pp. 690–705, 2022.
[9] 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.
[10] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 769–776. IEEE, 2017.
[11] Y. Wu, F. Zhang, and Y. Fu, "Real-time robotic multigrasp detection using anchor-free fully convolutional grasp detector," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 12, pp. 13 171–13 181, 2021.
[12] 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)*, pp. 598–605. IEEE, 2016.
[13] P. Schmidt, N. Vahrenkamp, M. Wächter, and T. Asfour, "Grasping of unknown objects using deep convolutional neural networks based on depth images," in *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 6831–6838. IEEE, 2018.
[14] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13 438–13 444. IEEE, 2021.
[15] H. Kasaei and M. Kasaei, "Mvgrasp: Real-time multi-view 3d object grasping in highly cluttered environments," *Robotics and Autonomous Systems*, vol. 160, p. 104313, 2023.
[16] 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.
[17] S. D'Avella, A. M. Sundaram, W. Friedl, P. Tripicchio, and M. A. Roa, "Multimodal grasp planner for hybrid grippers in cluttered scenes," *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 2030–2037, 2023.
[18] 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.
[19] 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*, pp. 11 444–11 453, 2020.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[21] 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*, pp. 15 964–15 973, 2021.
[22] V.-D. Nguyen, "Constructing force-closure grasps," *The International Journal of Robotics Research*, vol. 7, no. 3, pp. 3–16, 1988.
[23] J. Mahler, M. Matl, X. Liu, A. Li, D. 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," in *2018 IEEE International Conference on robotics and automation (ICRA)*, pp. 5620–5627. IEEE, 2018.
[24] 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.
[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[26] 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)*, pp. 13 459–13 466. IEEE, 2021.
[27] Z. Liu, Z. Chen, S. Xie, and W.-S. Zheng, "Transgrasp: A multi-scale hierarchical point transformer for 7-dof grasp detection," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 1533–1539. IEEE, 2022.
[28] R. Qin, H. Ma, B. Gao, and D. Huang, "Rgb-d grasp detection via depth guided learning with cross-modal attention," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8003–8009. IEEE, 2023.
[29] H. Ma and D. Huang, "Towards scale balanced 6-dof grasp detection in cluttered scenes," in *Conference on robot learning*, pp. 2004–2013. PMLR, 2023.
[30] S. Chen, W. Tang, P. Xie, W. Yang, and G. Wang, "Efficient heatmap-guided 6-dof grasp detection in cluttered scenes," *IEEE Robotics and Automation Letters*, 2023.
[31] J. Cai, H. Cheng, Z. Zhang, and J. Su, "Metagrasp: Data efficient grasping by affordance interpreter network," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 4960–4966. IEEE, 2019.
[32] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.
[33] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2016.

Wenshuo Wang received the B.Eng. degree in automation from the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China in 2021. He is currently working toward the Ph.D. degree at the Department of Mechanical Engineering, National University of Singapore, Singapore. His research interests include computer vision, robotic manipulation.

Haiyue Zhu received the B.Eng. degree in automation from the School of Electrical Engineering and Automation and the B.Mgt. degree in business administration from the College of Management and Economics, Tianjin University, Tianjin, China, in 2010, and the M.Sc. and Ph.D. degrees in electrical engineering from the National University of Singapore, Singapore, in 2013 and 2017, respectively. He is currently a Senior Scientist with the Adaptive Robotics and Mechatronics Group, Singapore Institute of Manufacturing Technology (SIMTech), Agency for Science, Technology and Research (A*STAR), Singapore. His research interests include intelligent mechatronics and robotic systems, smart robot grasping and manipulation, etc.



Marcelo H. Ang Jr received the B.Sc. degrees (CumLaude) in Mechanical Engineering and Industrial Management Engineering from the De La Salle University, Manila, Philippines, in 1981; the M.Sc. degree in Mechanical Engineering from the University of Hawaii at Manoa, Honolulu, Hawaii, in 1985; and the M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Rochester, Rochester, New York, in 1986 and 1988, respectively. He is currently a professor in the Department of Mechanical Engineering at the National University of Singapore. He is also the Director of the Advanced Robotics Centre. His research interests span the areas of robotics, mechatronics, and applications of intelligent systems methodologies. He teaches both at the graduate and undergraduate levels in the following areas: robotics; creativity and innovation, and Engineering Mathematics.