POLITECNICO DI TORINO
Repository ISTITUZIONALE

MultiGraspNet: A Multitask 3D Vision Model for Multi-Gripper Robotic Grasping

*Original*
MultiGraspNet: A Multitask 3D Vision Model for Multi-Gripper Robotic Grasping / Ortuno Chanelo, S., Rabino, P., Civitelli, E., Tommasi, T., Camoriano, R.. - In: IEEE ROBOTICS AND AUTOMATION LETTERS. - ISSN 2377-3766. - 11:9(2026), pp. 10919-10926. [10.1109/LRA.2026.3719235]

*Availability:*
This version is available at: 11583/3015433 since: 2026-09-14T10:26:30Z

*Publisher:*
Institute of Electrical and Electronics Engineers

*Published*
DOI:10.1109/LRA.2026.3719235

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

(Article begins on next page)

06 October 2026

# MultiGraspNet: A Multitask 3D Vision Model for Multi-Gripper Robotic Grasping

Stephany Ortuno-Chanelo, Paolo Rabino, Enrico Civitelli, Tatiana Tommasi, and Raffaello Camoriano, *Member, IEEE*

***Abstract*—Vision-based models for robotic grasping automate critical, repetitive, and draining industrial tasks. Existing approaches are typically limited in two ways: they either target a single gripper and are potentially applied on costly dual-arm setups, or rely on custom hybrid grippers that require ad-hoc learning procedures with logic that cannot be transferred across tasks, restricting their general applicability. In this work, we present *MultiGraspNet*, a novel multitask 3D deep learning method that predicts feasible poses simultaneously for parallel and vacuum grippers within a unified framework, enabling a single robot to handle both end-effectors. The model is trained on the richly annotated GraspNet-1Billion and SuctionNet-1Billion datasets, which have been aligned for the purpose, and generates graspability masks quantifying the suitability of each scene point for successful grasps. By sharing early-stage features while maintaining gripper-specific refiners, MultiGraspNet effectively leverages complementary information across grasping modalities. This design preserves a compact architectural footprint of only 15.75 M parameters and enables fast inference on a single GPU, enhancing adaptability and efficiency in cluttered scenes. We characterize MultiGraspNet's performance with an extensive experimental analysis, demonstrating its competitiveness with single-task models on relevant benchmarks while reducing computational cost. Moreover, real-world experiments on a single-arm multi-gripper robotic setup show that our approach outperforms normalization-based multi-gripper approaches.**

***Index Terms*—Robotic grasping, multi-gripper grasping, 3D perception, multitask learning.**

## I. Introduction

In RECENT years, the demand for robotic systems in industrial and warehouse environments has grown significantly, expanding toward increasingly diverse and complex tasks. This trend is driven by the need to improve operational efficiency and working conditions, while reducing safety risks for human operators. A central challenge in this context is the development of robotic systems capable of handling a wide variety of items that differ in shape, size, weight, and material properties. In response, research in autonomous grasping has led to increasingly sophisticated solutions that integrate computer vision, machine learning, and robotics to perceive, localize, and manipulate objects in cluttered environments.

Classical robotic grasping systems employ a single gripper, most commonly either a parallel gripper or a vacuum gripper. The former relies on frictional contacts and is best suited to rigid objects that provide sufficient friction and stable contact geometry for force closure [1], [2]. The latter is most effective for objects with smooth, low-porosity surfaces that can form a seal [3], [4], [5]. Although using one end-effector is cost-effective and simplifies grasp planning, performance is limited by the mechanical and geometric properties of the selected gripper. Thus, single-gripper robots often struggle to reliably grasp a diverse range of objects, reducing their adaptability and effectiveness. These limitations have been addressed through the development of custom hybrid grippers combining suction cups [6], compliant fingers [7], or sensor-integrated fingertips [8], [9]. Such grippers are typically tailored to specific object characteristics, achieving high performance in targeted applications, but exhibiting limited adaptability when confronted with novel items or unanticipated configurations, which hinders generalization.

Recently designed systems overcome these issues by using different grippers with multiple robotic arms, working collaboratively to handle a broader range of objects [10], [11]. However, they present increased spatial requirements, financial costs, and overall complexity. Furthermore, despite improving grasping capabilities, these systems typically rely on separate, gripper-specific models, which can lead to suboptimal learning. This setup is computationally inefficient during both training and prediction, since each model is trained independently. Besides, the learned representations remain isolated, preventing knowledge transfer and limiting generalization to unseen scenarios. In this context, multi-gripper methods for a single robotic arm represent an underexplored, yet promising direction that can strike the balance between cost and performance. Crucially, the challenge goes beyond designing and controlling a robot capable of switching among multiple end-effectors, and calls for learning a multi-objective shared model within a single framework. Such model should build a unified internal representation that captures geometric features relevant to different grasping modalities.

For this purpose, we introduce our *MultiGraspNet*, a novel 3D vision-based multitask deep learning architecture that predicts grasping poses for two different grippers simultaneously. Overcoming the need for multiple specialized models, MultiGraspNet leverages both early-stage shared representations and specialized pose refiners to extract tailored features for each gripper resulting in improved grasping success rate and a unified learning pipeline. Our approach, outlined in Fig. 1, allows a single robotic arm to utilize two distinct grippers: a parallel-jaw gripper and a vacuum gripper. This configuration enhances the versatility of the

Received 30 January 2026; accepted 11 June 2026. Date of publication 31 July 2026; date of current version 10 August 2026. This article was recommended for publication by Associate Editor S. Foix and Editor J. Borràs Sol upon evaluation of the reviewers' comments. This work was supported in part by Comau S.p.A. and in part by European Union Next-GenerationEU under FAIR - Future Artificial Intelligence Research project (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, under Grant PE00000013). *(Corresponding author: Stephany Ortuno-Chanelo.)*

Stephany Ortuno-Chanelo, Paolo Rabino, and Tatiana Tommasi are with the VANDAL Laboratory, Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy (e-mail: stephany.ortuno@polito.it).

Enrico Civitelli is with Comau S.p.A., Advanced Automation Solutions, 10095 Grugliasco, Italy.

Raffaello Camoriano is with the VANDAL Laboratory, Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy, and also with Istituto Italiano di Tecnologia, 16163 Genoa, Italy.

Project page: https://vandal-lab.github.io/multigraspnet-project.

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2026.3719235, provided by the authors.

Digital Object Identifier 10.1109/LRA.2026.3719235

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

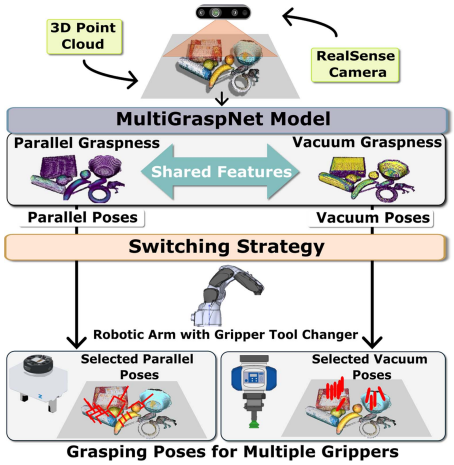


Fig. 1. Our proposed multi-gripper approach processes a 3D scene point cloud and learns a shared representation to jointly predict grasping quality scores for parallel and vacuum grippers based on geometric cues. The resulting graspness maps are further refined to produce the final grasp poses for both grippers. Finally, a switching strategy module selects the optimal gripper type for execution without altering the predicted poses.

robot, allowing it to handle a significantly broader range of objects while reducing hardware and computational costs.

*To summarize, our main contributions are:*

- We introduce MultiGraspNet, an efficient multitask 3D deep learning model capable of generating multiple candidate poses for vacuum and parallel grippers simultaneously from 3D visual input with a compact 15.75 M parameter footprint.
- We introduce a learning-based switching strategy that dynamically selects the most suitable gripper, effectively resolving the final execution choice between modality-specific grasp predictions.
- We evaluate our system with an extensive experimental analysis including benchmark comparisons, design-choice and real-world grasping experiments on known and novel objects in an industrial scenario.

Our results demonstrate MultiGraspNet's advantage over single-task baselines and highlight the potential of multitask learning for scalable and robust multi-gripper robotic manipulation, paving the way for future research on unified grasping models and more adaptive robotic systems.

## II. Related Works

### A. Suction Grasping

Suction-based grasping is a particularly effective solution in cluttered environments due to its simplicity. Early approaches to suction grasp prediction primarily relied on convolutional neural networks (CNNs) to infer grasp poses from RGB or depth images, while alternative methods exploited purely geometric reasoning on point clouds: [12] proposed a hybrid ResNet-based U-Net architecture for dense suction grasp prediction. SuctionNet-1Billion (S1B) [13] further advanced the field through a two-stage framework that integrates physics-based grasp evaluation and provides a large-scale benchmark for online evaluation.

Concerning 3D perception–based approaches, the Grasp Quality CNN (GQ-CNN) [3] method trained on the synthetic Dex-Net 3.0 dataset demonstrated strong generalization capabilities on novel objects. More recently, SGNet [14] showed how to predict suction grasp parameters and object centers directly from point cloud data, while UISN [4] proposed a multi-stage pipeline that focuses on unseen objects by using Unseen Object Instance Segmentation (UOIS) [15].

Despite their effectiveness, these methods tend to perform best on relatively regular and flat objects, while being less reliable on highly textured or deformable objects, and on surfaces that hinder the formation of an airtight suction seal.

### B. Parallel Grasping

Early approaches for parallel grasp detection introduced the grasping rectangle representation [16]. Building on this idea, [17] proposed a cascaded two-network system that prunes unlikely grasps to balance speed and robustness. Subsequent works introduced region-based detectors combining global and local candidates to improve robustness and precision [18]. More recently, RGBD-Grasp [19] decomposed parallel grasp prediction into two sub-tasks: inferring position and orientation, followed by grasp width and depth.

Among 3D perception-based methods, Dex-Net 2.0 [1] introduced a large-scale synthetic set of point clouds and grasping poses, with a first version of the GQ-CNN model that predicted parallel grasp success probabilities. Recognizing the need for larger datasets and better evaluation systems, GraspNet-1Billion [20] provided a massive data collection from the real world with annotations obtained by analytic computation in simulation. To improve grasp candidate selection, GSNet [21] introduced the concept of *graspness*, a geometry-driven quality measure, and developed a fast end-to-end network integrating it. The authors of [22] introduced an end-to-end deep learning approach leveraging differentiable sampling. More recently, EconomicGrasp (EFG) [2] tackled the challenge of costly supervision by efficiently selecting labels with low ambiguity. GenGrasp [23], instead, explicitly modeled grasping physical priors to enhance novel object manipulation.

Although parallel-gripper approaches are well-suited for handling complex objects, they can face difficulties with large or flat objects of limited height. In such cases, incorporating an end-effector with complementary capabilities, such as a vacuum gripper, can enhance effectiveness.

### C. Multi-gripper Grasping

To overcome the limitations of each end-effector, several studies focused on the adoption of multi-functional grippers. For instance, Carman [24] won the Amazon Robotics Challenge in 2017 with a custom grasping tool that included suction and parallel grippers. Their model relied on a semantic segmentation network to predict grasping poses, followed by heuristic grasp synthesis. In the same challenge, [6] presented an approach for predicting affordance maps over four grasping primitives, enabling the use of a custom gripper that combines suction and parallel grasping. Instead, [10] proposed a robot system able to use two gripper combinations via a switching strategy. It exploited a Single Shot MultiBox Detector (SSD) network that includes objectness prediction, and the choice of the gripper is based on the sparseness of the identified object items. In [11], a switching strategy was introduced for suction and multi-fingered grippers to perform random bin picking and assembly applications. Even in this case the learning component was limited to object recognition and localization.

Dex-Net 4.0 [25] trained separate GQ-CNNs to predict the probability of grasp success for parallel and suction grippers given a point cloud, rather than pursuing a unified model. In contrast, Corner-Grasp [26] introduced a multitask learning approach that infers 2D grasping poses for multiple grippers and predicts surface material to select the grasping modality.

While prior work on multi-gripper grasping began exploring the benefits of multitask learning for jointly modeling multiple grasping modalities, existing approaches are primarily validated on 2D data in synthetic environments or are constrained by custom hardware.

Building on this emerging direction, MultiGraspNet introduces a multitask framework that operates in 3D on real-world data, effectively unifying distinct grasping modalities by leveraging feature sharing within a single and efficient architecture.

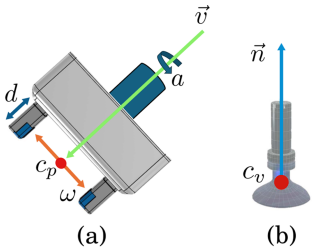


Fig. 2. Schematic illustration of the parallel-jaw gripper (a) and vacuum gripper (b) poses. The former includes the central point $c_p$, the depth $d$, the width $\omega$, as well as the approach direction $v$ and angle $a$. The vacuum grasp pose includes the central point $c_v$ and the normal to the point $n$.

## III. Method

### A. Task Definition

We introduce a unified multitask deep learning framework that enables a single model to predict grasping poses for two heterogeneous end-effectors, as illustrated in Fig. 2. Our approach is designed to increase the system's versatility by leveraging the advantages of each gripper within a single framework capable of learning features relevant to both tasks.

Given an input scene $\{p_i \in \mathbb{R}^3\}_{i=1}^{N}$, formally represented as a 3D point cloud of size $N$, the goal of our model is to predict candidate grasping poses for two gripper types simultaneously. Following the conventions established in [2], we define the 3D grasping pose for the parallel-jaw gripper as:

$$G_p = [c_p, v, a, \omega, d, s_p], \tag{1}$$

where $c_p$ indicates the central point of the grasp in the 3D space, $v$ is the approaching direction vector and $a$ [°] represents the angle of the 2D in-plane rotation. The remaining three components $d, \omega$, and $s_p$ are respectively the depth [m], which is the distance from the grasp point to the origin of the gripper frame, the width [m], that describes the distance between the fingers needed to grasp the object, and the grasp score which indicates the quality of the pose.

For the vacuum-based gripper, the grasp pose is represented as in [13]:

$$G_v = [c_v, n, s_v], \tag{2}$$

where $c_v$ is the central grasp point in the 3D space, $n$ is an estimate of the 3D surface normal vector in that point, and $s_v$ is the vacuum score indicating the quality of the pose.

### B. Dataset With Multi-gripper Graspness Map

To obtain a dataset with multi-gripper annotations, we align the GraspNet-1Billion [20] and SuctionNet-1Billion [12] datasets, which share a common set of scenes. We define a *graspness score map* that encodes the suitability of scene regions for grasping across two end-effectors, enabling multi-task training and fair comparison with single-task methods.

For the parallel gripper, the map is computed by starting from the provided friction coefficient and using the look-ahead search strategy for grasp success introduced in [21]. In this letter, this strategy is extended to the vacuum task over single objects by considering the provided seal coefficient and filtering out unsuccessful grasps (seal coeff $< 0.004$). Then, the graspness map is obtained at the scene level by using the 6D pose for each object and filtering out the poses with collisions. Finally, the map is projected onto the original point cloud.

To exclude non-graspable locations, points outside the operational zone (table depth $< 0$) are pruned, together with background points not belonging to any object instance using the *objectness* score $o$. Then, a nearest neighbor search associates each scene point with its closest grasp properties. The per-point vacuum graspness map is rescaled to the [0, 1] range, after which low-quality grasp regions are filtered out (scores below 0.1).

Overall, a supervised multi-gripper graspness map with the needed $r_p$ and $r_v$ values is constructed, derived from physically accurate grasping poses and accounting for the force-closure coefficient in the case of the parallel gripper and the seal coefficient for the vacuum gripper. This formulation enables the model to learn meaningful graspability patterns grounded in physically valid grasp configurations.

### C. MultiGraspNet

Our newly designed multi-gripper model is presented in Fig. 3. The rest of this section describes its main components.

*1) Backbone Network:* Similar to previous approaches [2], [21], we use a 3D U-Net as the main backbone, which is based on MinkowskiEngine [27]. The backbone encodes geometric scene information into per-point vectors, providing as output a feature tensor of shape $[N, 512]$.

*2) MultiGrasp Map Predictor:* This module consists of three heads that elaborate in parallel on the features to output, respectively, the objectness score $\hat{o}$, the graspness scores for the parallel gripper $\hat{r}_p$, and the graspness scores for the vacuum gripper $\hat{r}_v$. Every head is defined by a 1D convolutional layer that independently predicts a score for each point, obtaining a $[N, 3]$-dimensional output that can be visualized onto the scene as the graspness and objectness maps.

During training, the objectness head is optimized using a standard cross-entropy loss $\mathcal{L}_{obj}$ as in [2], while both graspness heads are trained using a weighted binary cross entropy loss:

$$\mathcal{L}_{graspness} = -\frac{1}{N}\sum_{i=1}^{N}[\lambda \cdot r_i \log(\hat{r}_i) + (1 - r_i)\log(1 - \hat{r}_i)]. \tag{3}$$

For the parallel graspness loss $\mathcal{L}^{par}_{graspness}$, we set the positive samples weight $\lambda = 10$ to account for graspable points sparsity, while $\lambda = 1$ for the vacuum graspness loss $\mathcal{L}^{vac}_{graspness}$.

*3) MultiGrasp Sampling:* Combining the task-specific graspness maps with objectness enables the identification of object points with a high probability of successful grasping. We formalize this via per-point score pair multiplication (denoted by $\otimes$ in Fig. 3), followed by thresholding ($t_p$=$t_v$=0.1) to discard low-scoring points. To ensure maximum spatial coverage, the points are sub-selected via Farthest Point Sampling (FPS). Finally, 1024 seed points for each gripper ($M_p$, $M_v$), paired with their 512-dimensional backbone features, are passed to the respective parallel and vacuum pose refiners.

*4) Parallel Pose Refiner:* Following [20] and [21], the $M_p$ seeds are fed to *ViewNet* and *Cylinder Grouping* to determine approaching vectors $v$. The training process optimizes a Smooth L1 loss for view ($\mathcal{L}_{view}$) and width prediction ($\mathcal{L}_{width}$), while a cross-entropy loss is adopted for the in-plane angle ($\mathcal{L}_{angle}$) and depth ($\mathcal{L}_{depth}$). Meanwhile, the grasp score $s_p$ estimation is formulated as a classification task and supervised via the cross-entropy loss $\mathcal{L}_{score}$.

*5) Vacuum Pose Refiner:* The vacuum pose prediction follows a different strategy due to the simpler nature of the pose. The refiner does not need further losses and is only adopted at inference time (i.e., $s_v = \hat{r}_v$). For each of the points in $M_v$, the surface normal vector $n$ is estimated by looking at the geometry of its nearby neighbors. Using an off-the-shelf normal estimation method [28], a small radius is used to find the closest points around each seed, and the normal vector is computed by analyzing the local shape through covariance to fit the best local surface. These normals complete the vacuum poses directly from point geometry without requiring additional view selection or grouping operations.

Overall, the design of our multi-gripper model combines the advantages of feature sharing and task-specific refiners to handle the unique characteristics of different end-effectors, enabling multi-grasp predictions within a single unified framework, avoiding the use of costly separate architectures.

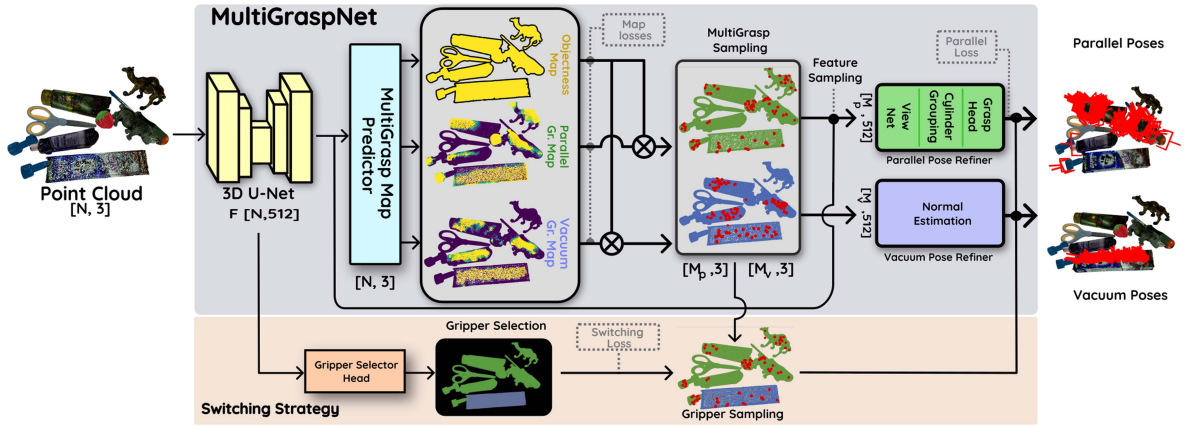


Fig. 3. The network takes as input a 3D point cloud. Then the Minkowski-based backbone extracts geometric features and these are processed by a multi-branch grasp prediction head. Gripper-specific refinement modules are applied to generate the multi-gripper grasping poses. Finally, a Switching Strategy module predicts a mask to filter these poses by the optimal gripper, without modifying the underlying MultiGraspNet predictions.

The entire network is trained end-to-end. The total loss $\mathcal{L}_{total}$ for MultiGraspNet is formulated by combining the individual task-specific losses as:

$$\mathcal{L}_{total} = \mathcal{L}_{obj} + \alpha\mathcal{L}^{vac}_{graspness} + \beta\mathcal{L}^{par}_{graspness} + \gamma\mathcal{L}_{view} + \mathcal{L}_{angle} + \mathcal{L}_{depth} + \mathcal{L}_{score} + \delta\mathcal{L}_{width}, \quad (4)$$

where ad-hoc weights are introduced and calibrated to maintain comparable scales across the loss terms. We set the view selection weight as $\gamma = 100$, while the weights for the width prediction, parallel, and vacuum grippers are respectively set as $\delta = \alpha = \beta = 10$.

*6) Switching Strategy:* A key challenge in multi-gripper grasp prediction is that the grasp quality scores produced for different end-effectors are not directly comparable, as $s_p$ and $s_v$ are derived from fundamentally different physical principles. Thus, a naïve comparison between these scores may lead to ambiguous or suboptimal gripper selection when both types yield high confidence on the same object area. To address this issue, we introduce a dedicated *Gripper Selector Head* that determines the most suitable gripper type for each object based on its geometric properties. Concretely, we implement an additional head that operates on the pretrained 3D U-Net backbone features. This module predicts a per-point probability vector $\hat{y} \in (0,1)^3$, whose entries correspond to the background, parallel, or vacuum gripper. The one-hot encoded ground-truth vector $y \in \{0,1\}^3$ for this head is annotated based on geometric considerations. Specifically, labels are assigned by expert human operators primarily evaluating local surface flatness.[1] During training, this head is optimized using a multi-class cross-entropy loss:

$$\mathcal{L}_{switching} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{3} y_{k,i}\log(\hat{y}_{k,i}). \quad (5)$$

At inference time, the predicted vector is used to filter the grasp candidates generated by MultiGraspNet. In particular, for each object, only the grasp poses corresponding to the selected gripper type are retained.

This enforces a consistent assignment of a single gripper per object, preventing conflicting grasp proposals.

### D. Implementation Details

Our approach addresses two tasks corresponding to two end-effectors (vacuum and parallel). To efficiently integrate these tasks, we adopt the Gradient Surgery for Multitask Learning (PCGrad) technique [29] during training, which mitigates conflicting gradient updates across tasks.

For training, we use the Adam optimizer with an initial learning rate of $5 \times 10^{-4}$ and a cosine decay scheduler. The batch size is set to 12, and training runs for 22 epochs. All experiments are conducted using PyTorch on a single NVIDIA GeForce RTX 5090 GPU.

[1]G1B [20] object IDs assigned to vacuum: $\{1, 2, 7, 36, 40, 47, 59, 69\}$. All the remaining ones are assigned to parallel.

## IV. Experiments

### A. Experimental Protocol and Metrics

The defined dual-gripper dataset comprises 190 scenes, of which 100 are used for training and 90 for testing. Following [13], [20], we divide the test set into three groups: seen/similar/novel. This enables us to evaluate how different methods generalize in increasingly complex scenarios. The grasp prediction modules generate multiple grasp candidates for each scene. Therefore, by following standard practice [13], [20], we adopt Precision@$k$ as our evaluation metric, which measures precision among the top-k ranked grasps. We denote by $AP_{\mu_p}$ and $AP_{\mu_v}$ the Average Precision@$k$ for the parallel and vacuum tasks respectively, with $k$ ranging from 1 to 50. The subscripts $\mu_p$ and $\mu_v$ refer to specific friction (parallel) and seal (vacuum) coefficients. Note that a high $\mu_p$ indicates that the gripper applies a large force, which generally leads to a higher-quality grasp. For the vacuum gripper, a grasp is considered positive only when the seal coefficient exceeds the value indicated by the subscript. Therefore, a high $\mu_v$ corresponds to a more selective and stricter success criterion. We also report the overall AP as the average result with $\mu_p$ ranging from 0.2 to 1.0 and $\mu_v$ ranging from 0.2 to 0.8, both with an interval of 0.2.

### B. Benchmark Comparison

In this first analysis, we are interested in validating the core functioning of our model, so we report the performance of its single-task variants, either vacuum (*MultiGraspNet-vac.*) or parallel (*MultiGraspNet-par.*), obtained by deactivating the objective of the other gripper during training. The multitask version optimizes the loss in (4), thus it does not include the switching strategy. We compare the results of our MultiGraspNet with well-established single-gripper approaches, including S1B [13], Dex-Net3.0 [3] for the vacuum, and G1B [20], GSNet [18], EFG [2] for the parallel end-effector.

The upper section of Table I presents the vacuum grasping results. For $AP_{0.8}$, our vacuum-specific variant MultiGraspNet-vac. consistently outperforms competing methods across all object groups. This

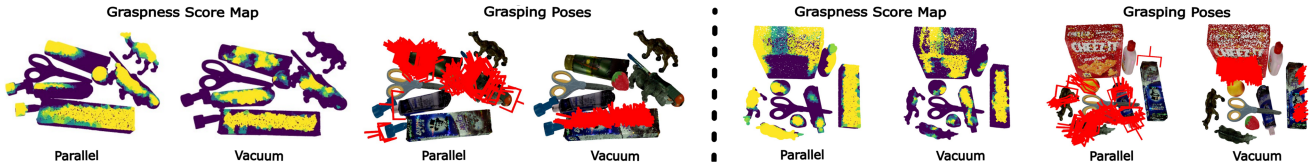


Fig. 4. Qualitative results of our MultiGraspNet for two different scenes. Our model outputs tailored graspness score maps for each end-effector, which then guide the prediction of the grasping poses in an end-to-end framework. The shown grasping poses are the top 100 obtained for each gripper after the filtering provided by the switching strategy module.

TABLE I
SINGLE-TASK RESULTS ON THE REALSENSE DATA OF THE ALIGNED DUAL-GRIPPER DATASET FROM SUCTIONNET-1BILLION (VACUUM) AND GRASPNET-1BILLION (PARALLEL). TOP RESULTS ARE IN BOLD, WHILE THE SECOND BEST ARE UNDERLINED.

| **Vacuum** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Seen | | | Similar | | | Novel | | |
| Method | $AP$ | $AP_{0.8}$ | $AP_{0.4}$ | $AP$ | $AP_{0.8}$ | $AP_{0.4}$ | $AP$ | $AP_{0.8}$ | $AP_{0.4}$ |
| DexNet3.0 [3] | 15.50 | 1.53 | 20.22 | 18.92 | 2.62 | 24.51 | 2.62 | 0.35 | 5.32 |
| S1B [13] | **28.31** | 3.41 | **38.56** | 26.64 | 3.42 | 35.34 | **8.23** | 0.35 | **10.29** |
| MultiGraspNet-vac. | 28.26 | 3.60 | 36.39 | **31.37** | **3.88** | 41.63 | 8.07 | **0.48** | 10.03 |
| MultiGraspNet | 28.07 | **3.62** | 35.81 | 31.22 | 3.53 | **41.65** | 7.38 | 0.46 | 8.92 |
| **Parallel** | | | | | | | | | |
| | Seen | | | Similar | | | Novel | | |
| Method | $AP$ | $AP_{0.8}$ | $AP_{0.4}$ | $AP$ | $AP_{0.8}$ | $AP_{0.4}$ | $AP$ | $AP_{0.8}$ | $AP_{0.4}$ |
| G1B [20] | 27.56 | 33.43 | 16.95 | 26.11 | 34.18 | 14.23 | 10.55 | 11.25 | 3.98 |
| GSNet [21] | 67.12 | 78.46 | 60.90 | 54.81 | 66.72 | 46.17 | 24.31 | 30.52 | **14.32** |
| EFG [2] | 68.21 | 79.60 | 63.54 | **61.19** | **73.60** | 53.77 | 25.48 | 31.46 | 13.85 |
| MultiGraspNet-par. | 67.70 | 78.97 | 62.67 | 60.61 | 72.51 | **54.01** | 25.45 | 31.36 | 13.60 |
| MultiGraspNet | **68.37** | **79.61** | **64.02** | 60.28 | 72.54 | 52.87 | **25.53** | **31.67** | 13.66 |

TABLE II
SMALL-DATA REGIME ANALYSIS

| | | Number of scenes | | | | | |
|---|---|---|---|---|---|---|---|
| **Metric** | **Model** | **1** | **2** | **3** | **4** | **5** | **6** |
| $\overline{AP_v}$ | MultiGraspNet-vac. | 15.19 | 16.66 | 17.59 | 18.26 | 18.17 | 18.85 |
| | MultiGraspNet | 15.14 | 17.17 | 18.12 | 17.80 | 18.71 | 19.15 |
| $\overline{AP_p}$ | MultiGraspNet-par. | 15.30 | 20.73 | 25.32 | 29.33 | 32.74 | 33.59 |
| | MultiGraspNet | 14.48 | 20.38 | 24.16 | 28.95 | 32.18 | 33.93 |

advantage is particularly pronounced for similar objects, where improvements are also observed for $AP_{0.4}$ and $AP$. The full multitask version MultiGraspNet preserves the gain in most of the considered cases, with minor positive or negative fluctuations.

The bottom portion of the table reports the results for parallel grasping. All competing methods in this setting are 3D point-cloud–based and have model capacities comparable to ours. While the single-task variant MultiGraspNet-par. achieves slightly lower performance than the best EFG competitor with a gap of less than one $AP$ point, the full multitask MultiGraspNet attains the best results on both the seen and novel object groups.

We also consider a sub-sampled training set consisting of one to six scenes. We report the average AP across splits (seen, similar, novel) for vacuum $\overline{AP_v}$ and parallel $\overline{AP_p}$ tasks in Table II. The results confirm that MultiGraspNet maintains competitive performance to the single-task variants across both modalities even in small data regimes.

Fig. 4 presents qualitative results obtained by MultiGraspNet. Our model successfully generates feasible grasps for both grippers in diverse scenes. Note that the predicted vacuum poses primarily concentrate on flat and smooth surfaces, such as boxes or planar regions (e.g., Cheez it box). Instead, parallel grasp predictions focus on smaller and more complex objects (e.g., screws). This complementary behavior highlights MultiGraspNet's ability to adapt to the distinct properties of different grippers.

### C. Exploring Design Variants

Here, we provide a detailed analysis of how key design choices of MultiGraspNet contribute to its performance. We empirically show that employing gradient surgery in multitask training and excluding color features yields overall quantitative advantages across tasks.

*1) Gradient Surgery for Multitask Learning:* Multitask learning often suffers from optimization challenges due to gradient interference between tasks, which can lead to imbalanced training and cause certain tasks to dominate the learning process. Here, we compare the performance of MultiGraspNet trained with the Adam optimizer with and without (w/o) PCGrad. As shown in Table III, the model without PCGrad achieves noticeable improvements in parallel/seen and vacuum/novel, albeit at the cost of decreased performance across other groups. To provide a single summary metric for task performance, we average the $AP$ values across the three categories (seen, similar, and novel) for each gripper, and we name it $\overline{AP}$. The performance of our model using just the Adam optimizer results in $\overline{AP}_p = 51.39$ for parallel and $\overline{AP}_v = 21.29$ for vacuum. When incorporating PCGrad, MultiGraspNet yields $\overline{AP}_p = 51.39$ for parallel and $\overline{AP}_v = 22.22$ for vacuum. We conclude that PCGrad leads to a more balanced and consistent performance across grasping modalities and splits. Thus, we always train our model with PCGrad.

*2) Adding RGB Features:* To test whether additional color information could improve scene understanding, we extend our model to incorporate 2D RGB features extracted from a pretrained ResNet-18 backbone. They are concatenated with the 3D features obtained from the point cloud, allowing the network to leverage both geometric and appearance-based cues. As shown in Table III, MultiGraspNet with (w) RGB features underperforms across tasks and splits with a small advantage only on novel objects for the vacuum gripper. These findings indicate that, within our setup, adding color information does not provide a reliable benefit and may even introduce unnecessary complexity without improving grasping performance. Thus, we decided not include RGB data for training our model.

*3) Corner-Grasp Backbone:* To further assess the impact of the feature extraction module, we conduct an additional comparison with the Corner-Grasp [26]. Although the original implementation is not publicly available, we replicate its backbone by adopting a ResNet101+FPN configuration and integrating it into our framework as a replacement for the 3D U-Net backbone. This modification allows us to isolate the contribution of the backbone while keeping the remaining components of our pipeline unchanged. The results in Table III highlight the effectiveness of our original design, showing that the 3D U-Net backbone provides more suitable geometric features for grasp prediction. Leading to improved performance particularly for the parallel gripper, which requires richer 3D representations to handle more complex geometries. Specifically, our model outperforms CornerGrasp [26] baseline across all categories, achieving a significant margin of over 2.0 AP points in the parallel gripper tasks and showing competitive results in the vacuum grasping modality.

TABLE III
ANALYSIS OF DESIGN VARIANTS FOR MULTIGRASPNET. RESULTS ON THE REALSENSE DATA OF THE ALIGNED DUAL-GRIPPER DATASET FROM SUCTIONNET-1BILLION (VACUUM) AND GRASPNET-1BILLION (PARALLEL). TOP RESULTS ARE IN BOLD, WHILE THE SECOND BEST ARE UNDERLINED.

| | **Parallel** | | | | | | | | | **Vacuum** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Seen | | | Similar | | | Novel | | | Seen | | | Similar | | | Novel | | |
| Model | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ | AP | $AP_{0.8}$ | $AP_{0.4}$ |
| Corner-Grasp [26] backbone | 66.36 | 77.16 | 61.88 | 45.32 | 55.39 | 35.92 | 21.17 | 26.47 | 11.33 | 26.74 | 3.31 | 34.91 | 28.96 | **3.69** | 35.92 | **9.34** | 0.43 | **11.33** |
| MultiGraspNet w/o PCGrad | **69.18** | **80.00** | **65.13** | 60.08 | **72.59** | 52.33 | 24.92 | 30.75 | 13.41 | 26.70 | 2.90 | 34.60 | 28.54 | 2.86 | 37.02 | 8.64 | **0.49** | 10.56 |
| MultiGraspNet w/ RGB | 67.09 | 78.19 | 62.07 | 58.49 | 70.60 | 50.73 | 24.10 | 29.80 | 12.86 | 27.22 | 3.01 | 35.22 | 30.06 | 3.40 | 39.35 | 7.57 | 0.41 | 9.40 |
| MultiGraspNet | 68.37 | 79.61 | 64.02 | **60.28** | 72.54 | **52.87** | **25.53** | **31.67** | **13.66** | **28.07** | **3.62** | **35.81** | **31.22** | 3.53 | **41.65** | 7.38 | 0.46 | 8.92 |

TABLE IV
MODEL EFFICIENCY COMPARISON

| | **# Parameters ↓** | **Size [MB] ↓** | **FPS ↑** | **Latency [ms] ↓** |
|---|---|---|---|---|
| S1B [13] | $58.75 \times 10^6$ | 224.12 | 12.48 | 80.09 |
| EFG [2] | $15.75 \times 10^6$ | 60.08 | 21.97 | 45.50 |
| S1B+EFG | $74.50 \times 10^6$ | 284.20 | 7.91 | 125.59 |
| CornerGrasp [26] | $108.97 \times 10^6$ | 415.70 | 19.50 | 51.27 |
| MultiGraspNet | $15.75 \times 10^6$ | 60.09 | 19.69 | 50.78 |

Fig. 5. Left: Our multi-gripper system with depth camera includes a parallel and vacuum gripper stored in dedicated stations (2). The automated tool changer (1) facilitates quick switching by connecting directly to the gripper and removing it from its station. Right: Real-world test objects.

## D. Efficiency Analysis

We quantitatively evaluate the computational efficiency of our approach by comparing it with existing methods in terms of model size, number of parameters, and runtime performance. As reported in Table IV, our model achieves a favorable trade-off between accuracy and efficiency, requiring only $15.75 \times 10^6$ parameters and 60.09 MB memory, which is significantly lower than both Corner-Grasp and the combined S1B+EFG setup. Besides, our method maintains a competitive runtime performance, achieving a latency of 50.78 ms on a single GPU. These results demonstrate that MultiGraspNet can effectively handle multi-gripper grasp prediction without incurring in substantial computational overhead, and is therefore well-suited for real-time robotic applications under resource constraints.

## E. Real-World Experiments

To conduct a comprehensive evaluation of our proposed approach, we perform real-world experiments in an industrial setting. Qualitative video demonstrations of the grasp executions are provided in the supplementary material. All experiments are executed using a 6-DoF industrial robotic arm, equipped with a RealSense camera for 3D perception, as shown in Fig. 5. The end-effector configuration consists of a parallel gripper and a vacuum gripper, coupled via an automated tool changer that enables their rapid switching during operation.

To quantify the performance and to understand whether failures are due to the quality of the grasping pose or perceptual limitations, we employ the following metrics:

- $R_{object}$: Ratio of the number of successfully cleared objects to the total number of objects.
- $R_{grasp}$ *(success rate):* Ratio of the number of successful grasps to the total number of grasps.
- $R_{mix}$: Ratio of the number of grasps (over the successfully cleared objects) to the number of successfully cleared objects.
- $R_{seen}$: Ratio of the number of detected objects to the total number of objects in the scene.

We adopt the same experimental setup as previous works [2], [12], [20]. For each trial, the system processes the input point cloud and predicts grasp candidates. The one with the highest score is executed. After each attempt, the system re-evaluates the scene and repeats the process. The cycle continues until all objects are successfully cleared from the table or the robot fails to grasp any object for three consecutive times.

To assess performance in diverse cluttered environments, we consider ten test scenes belonging to the following categories:

- *Seen objects:* Five scenes are populated by randomly arranging six items from the seen objects (Fig. 5). This setup, with a total of 30 object instances, ensures variability in shape, texture, and physical properties while including repetitions across scenes.
- *Unseen objects:* Five scenes are composed using five randomly selected items from completely new objects. As shown in Fig. 5, these items introduce novel shapes, colors, deformable materials and irregular surfaces to evaluate the model's generalization capabilities across 25 total objects.

To establish a baseline, we compare MultiGraspNet against two single-gripper methods, S1B [13] for vacuum and EFG [2] for parallel grasping in two different modalities.

*1) Single-Gripper:* For a direct comparison with single-gripper methods, we evaluate MultiGraspNet by isolating the vacuum and parallel predictions against their respective single-gripper counterparts. The results reported in Table V indicate that our approach successfully grasps more objects with fewer attempts than S1B [13], as reflected by the higher $R_{object} = 70\%$ and $R_{grasp} = 41.02\%$ for seen objects, and $R_{object} = 80\%$ and $R_{grasp} = 54.76\%$ for the unseen ones. For single-parallel grasping, we obtain competitive results compared with EFG [2] for the seen objects by picking one more item ($R_{object} = 96\%$) with fewer attempts ($R_{grasp} = 76.64\%$). For the unseen objects, our method grasps one more object, although requiring slightly more attempts.

*2) Multi-Gripper With Online Switching:* To analyze the performance of MultiGraspNet including autonomous gripper selection and switching, we compare our method using the proposed switching strategy (named *MultiGraspNet w/ SS*) with the following baselines:

- *S1B+EFG w/ z-score:* Combining S1B+EFG via z-score normalization;
- *S1B+EFG w/ SS:* A version of S1B+EFG utilizing a dedicated module analogous to our switching strategy;
- *MultiGraspNet w/ z-score:* Our method adapted with z-score normalization to ensure a fair evaluation.

TABLE V
REAL-WORLD EXPERIMENTS PERFORMANCE

| Method | Seen Objects | | | | Unseen Objects | | | |
|---|---|---|---|---|---|---|---|---|
| | $R_{object}\uparrow$ | $R_{grasp}\uparrow$ | $R_{mix}\downarrow$ | $R_{seen}\uparrow$ | $R_{object}\uparrow$ | $R_{grasp}\uparrow$ | $R_{mix}\downarrow$ | $R_{seen}\uparrow$ |
| **Single-gripper** | | | | | | | | |
| | **Vacuum** | | | | | | | |
| S1B [13] | 16/30 (53%) | 39.11% | **1.14** | 24/30 (80%) | 12/25 (48%) | 43.33% | 1.17 | 17/25 (68%) |
| MultiGraspNet | **21/30 (70%)** | **41.02%** | 1.34 | **30/30 (100%)** | **20/25 (80%)** | **54.76%** | **1.14** | **25/25 (100%)** |
| | **Parallel** | | | | | | | |
| EFG [2] | 28/30 (93%) | 66.03% | 1.45 | **30/30 (100%)** | 22/25 (88%) | **76.98%** | **1.14** | 24/25 (96%) |
| MultiGraspNet | **29/30 (96%)** | **76.64%** | **1.35** | 29/30 (96%) | **23/25 (92%)** | 70.48% | 1.16 | **25/25 (100%)** |
| **Multi-gripper with online switching** | | | | | | | | |
| S1B + EFG w/ z-score | 22/30 (73%) | 42.52% | **1.32** | 29/30 (96%) | 18/25 (72%) | 50.79% | 1.40 | 23/25 (92%) |
| S1B + EFG w/ SS | 23/30 (76.67%) | 49.65% | 1.48 | 30/30 (100%) | 16/25 (64%) | **80%** | 1.4 | 18/25 (72%) |
| MultiGraspNet w/ z-score | 25/30 (83.33%) | 57.23% | **1.32** | **30/30 (100%)** | 21/25 (84%) | 54.06% | 1.38 | **25/25 (100%)** |
| MultiGraspNet w/ SS | **29/30 (96%)** | **70.83%** | 1.35 | **30/30 (100%)** | **24/25 (96%)** | 74.29 % | **1.16** | **25/25 (100%)** |

Table V shows that S1B+EFG w/ z-score results in large performance reductions with respect to the single-task EFG baseline on both seen and unseen objects (e.g., $-20\%$ and $-16\%$ in terms of $R_{object}$, respectively). We observe a similar trend for MultiGraspNet w/ z-score, which successfully grasps markedly fewer objects than parallel-only MultiGraspNet. These results show that simple normalization yields lower performance than single-gripper systems, highlighting that appropriate gripper selection is critical for multi-gripper systems.

Finally, we evaluate MultiGraspNet and S1B+EFG using our learning-based switching strategy for both methods. We notice that S1B+EFG w/ SS slightly improves in terms of $R_{object}$ for seen objects, while it decreases for unseen ones. However, in both cases it still performs poorly with respect to single-gripper baselines. On the contrary, MultiGraspNet w/ SS achieves $R_{object} = 96\%$ on both seen and unseen objects, matching or surpassing all baseline results, including S1B+EFG w/ SS.

Overall, our experimental results show that incorporating shared features allows MultiGraspNet to achieve a more complete understanding of the scene compared to the isolated, single-task features of the baselines. This results in higher performance on downstream multi-gripper grasping tasks.

## V. CONCLUSION

We present MultiGraspNet, a multitask 3D deep learning model capable of predicting grasp poses for both parallel-jaw and vacuum grippers within a unified framework. By leveraging complementary information between grasping modalities, our model generates accurate grasp poses across a variety of cluttered scenes with diverse objects. Furthermore, our multi-gripper experiments with online switching highlight the advantages of our unified architecture, significantly outperforming traditional normalization-based multi-gripper approaches and combined single-task baselines in real-world scenarios. By integrating multiple modalities in a shared architecture, MultiGraspNet provides an accurate, efficient, and versatile solution for real-world robotic grasping. Future work will investigate the integration of a supervision-free switching strategy alongside a learning-based module to minimize the number of gripper switches while accounting for picking ordering and collisions.

## ACKNOWLEDGMENT

The authors would like to thank Giovanni Di Stefano, Simone Panicucci, Nicola Longo, Luca Di Ruscio, Luca Robbiano, and Simone Peirone for their valuable assistance. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

## REFERENCES

[1] J. Mahler et al., "Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Proc. 13th Conf. Robot. Sci. Syst.*, 2017.
[2] X.-M. Wu, J.-F. Cai, J.-J. Jiang, D. Zheng, Y.-L. Wei, and W.-S. Zheng, "An economic framework for 6-DoF grasp detection," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 357–375.
[3] 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 *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 5620–5627.
[4] R. Cao, B. Yang, Y. Li, C.-W. Fu, P.-A. Heng, and Y.-H. Liu, "Uncertainty-aware suction grasping for cluttered scenes," *IEEE Robot. Automat. Lett.*, vol. 9, no. 6, pp. 4934–4941, Jun. 2024.
[5] D. Ceschini, R. D. Cesare, E. Civitelli, and M. Indri, "Segmentation-based approach for a heuristic grasping procedure in multi-object scenes," in *Proc. IEEE 29th Int. Conf. Emerg. Technol. Factory Automat.*, 2024, pp. 1–4.
[6] A. Zeng et al., "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," *Int. J. Robot. Res.*, vol. 41, pp. 690–705, 2022.
[7] W. Zhu et al., "A soft-rigid hybrid gripper with lateral compliance and dexterous in-hand manipulation," *IEEE/ASME Trans. Mechatron.*, vol. 28, no. 1, pp. 104–115, Feb. 2023.
[8] M. Burgess and E. H. Adelson, "Grasp everything (GET): 1-DoF, 3-fingered gripper with tactile sensing for robust grasping," 2025, *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2025.
[9] Y. Hu, M. Li, S. Yang, X. Li, S. Liu, and M. Li, "Learning robust grasping strategy through tactile sensing and adaption skill," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2024, pp. 1197–1202.
[10] M. Fujita et al., "Bin-picking robot using a multi-gripper switching strategy based on object sparseness," in *Proc. IEEE 15th Int. Conf. Automat. Sci. Eng.*, 2019, pp. 1540–1547.
[11] A. S. Olesen, B. B. Gergaly, E. A. Ryberg, M. R. Thomsen, and D. Chrysostomou, "A collaborative robot cell for random bin-picking based on deep learning policies and a multi-gripper switching strategy," *Procedia Manuf.*, vol. 51, pp. 3–10, 2020.
[12] Q. Shao et al., "Suction grasp region prediction using self-supervised learning for object picking in dense clutter," in *Proc. IEEE 5th Int. Conf. Mechatronics Syst. Robot.*, 2019, pp. 7–12.
[13] H. Cao, H.-S. Fang, W. Liu, and C. Lu, "SuctionNet-1billion: A large-scale benchmark for suction grasping," *IEEE Robot. Automat. Lett.*, vol. 6, no. 4, pp. 8718–8725, Oct. 2021.
[14] D.-H. Zhai, S. Yu, Y. Guan, and Y. Xia, "SGNet: Robotic suction grasp detection with multiscale attention," *IEEE/ASME Trans. Mechatron.*, vol. 30, no. 6, pp. 6927–6938, Dec. 2025.
[15] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "Unseen object instance segmentation for robotic environments," *IEEE Trans. Robot.*, vol. 37, no. 5, pp. 1343–1359, Oct. 2021.
[16] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2011, pp. 3304–3311.
[17] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, no. 4-5, pp. 705–724, 2015.
[18] U. Asif, J. Tang, and S. Harrer, "Densely supervised grasp detector (DSGD)," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2019, pp. 8085–8093.
[19] 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 *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 13459–13466.
[20] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11444–11453.

[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 *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 15964–15973.

[22] A. Alliegro, M. Rudorfer, F. Frattin, A. Leonardis, and T. Tommasi, "End-to-end learning to grasp via sampling from object point clouds," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 9865–9872, Oct. 2022.

[23] H. Ma, M. Shi, B. Gao, and D. Huang, "Generalizing 6-DoF grasp detection via domain prior knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 18102–18111.

[24] D. Morrison et al., "Cartman: The low-cost cartesian manipulator that won the amazon robotics challenge," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 7757–7764.

[25] J. Mahler et al., "Learning ambidextrous robot grasping policies," *Sci. Robot.*, vol. 4, no. 26, pp. 1373–1379, 2019.

[26] Y. G. Son, S. Um, J. Hong, T. H. Bui, and H. R. Choi, "Corner-grasp: Multi-action grasp detection and active gripper adaptation for grasping in cluttered environments," 2025, *arXiv:2504.01861*.

[27] C. Choy, J. Gwak, and S. Savarese, "4D spatio-temporal convnets: Minkowski convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3075–3084.

[28] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3d data processing," 2018, *arXiv:1801.09847*.

[29] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 5824–5836.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement