# Statistical Stratification and Benchmarking of Robotic Grasping Performance

Brice Denoun[1,2], Miles Hansard[1], Beatriz León[2] and Lorenzo Jamone[1]

***Abstract*— Robotic grasping is fundamental to many real-world applications, and new approaches must be systematically evaluated. However, in most cases, the performance of a specific approach is assessed by simply counting the number of successful attempts in a given task, and this success rate is then compared to those of other solutions, without taking into account the random variability across different experiments (e.g. due to sensor noise, or variations in object placement). In order to address this issue, we classify the observed performance into qualitatively ordered outcomes, thereby stratifying the results. We then show how to analyse these results, in a statistical framework which accounts for the variability between experiments. The advantages of our approach are demonstrated in the practical comparison of four grasp planning algorithms. In particular, we show that the proposed approach allows us to carry out several distinct evaluations from a single set of experiments, without having to repeat the data collection process. We demonstrate that differences between the algorithms, which would not be apparent from overall success rates, can be identified and evaluated.**

## I. Introduction

Robotic grasping is an active research area which has been studied for several decades, owing to its importance in a wide range of industrial tasks [1]. The ongoing scientific contributions can be divided into the following three areas. First, advances in robot hardware have provided more reliable robot arms and anthropomorphic, soft or under-actuated manipulators [2]. Second, vision-based algorithms have been developed in order to extract useful information from the robot's environment, including object recognition, semantic segmentation [3], and pose detection [4]. Third, recent works have succeeded in providing robots with high-level skills, such as the ability to grasp objects autonomously [5].

Although newly published grasping algorithms are usually compared to the state-of-the-art, the underlying methodology used to quantify their performance tends to differ from one study to another [6], [7], [8]. Even the definition of a 'successful' grasp varies across the literature and often depends on the targeted use case [9]. In order to unify the evaluation of grasping and manipulation algorithms, a set of benchmark protocols and a standardised set of objects have been proposed [10], [11], [12]. Although these protocols offer experimental procedures and evaluation metrics, they do not provide a statistical methodology for performance comparison. A common approach is to compute an overall performance metric (typically the count of trials considered successful), and then to rank the methods accordingly [13], [14]. However, simple comparisons do not account for random variability across different experiments (e.g. noisy sensor data, error in object placement, etc.), which might lead to different conclusions if the experiments were to be repeated.

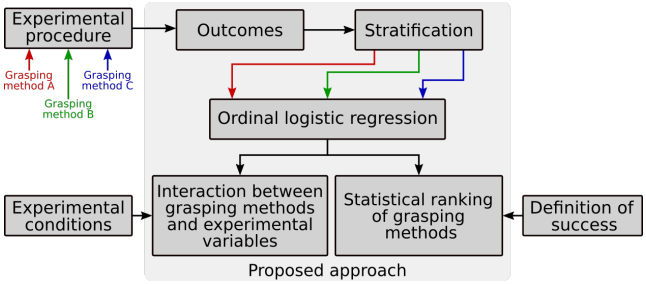



Fig. 1: Schematic diagram of our stratified approach for comparing robotic grasping algorithms. Unlike standard approaches, the concept of success is only introduced at the end of the comparison. This allows us to run a wider range of statistical analyses, based on a single set of experiments.

We illustrate this phenomenon through an empirical study and propose a generic *stratified* statistical approach for the analysis of grasping experiments in the presence of random effects. We show that using an ordered range of outcomes is beneficial for two reasons. First, it is possible to statistically compare the performance of grasping solutions according to different definitions of *success* from the same set of observed outcomes (see Figure 1). This allows for reporting more objective and robust results, regardless of the definition of 'success'. Second, the same statistical framework allows for more in-depth analysis of benchmark results considering multiple experimental variables that can influence the observed performance (e.g. choice of RGB-D sensors, end-effector, etc.). In this work, we propose a study to determine if the ranking of four grasp planning algorithms depends on the test object, as well as on its pose with respect to the depth camera.

The rest of the paper, which starts with a literature review in section II, is organised as follows. In section III, we present a study illustrating the shortcomings of comparing the success rate of grasp planning algorithms without accounting for the presence of random effects. Section IV formalises the definition of a stratification of outcomes, which is the basis of the statistical framework proposed in Section V. Finally, we demonstrate the benefits of our

*Work partially supported by the EPSRC UK (NCNR, EP/R02572X/1 and MAN[3], EP/S00453X/1) and by the Shadow Robot Company.

[1] ARQ (Advanced Robotics at Queen Mary), School of Electronic Engineering and Computer Science, Queen Mary University of London, UK {b.d.denoun, m.hansard, l.jamone}@qmul.ac.uk

[2] The Shadow Robot Company, London, United Kingdom

approach in Section VI.

## II. Related work

In the past two decades, several studies have examined the question of how to compare robotic systems, on similar tasks, in an unbiased way [15], [16]. For grasping and manipulation tasks, the YCB object set has been proposed [12] to establish a standard and reproducible set of physical objects that can be used across different tasks and studies. Although defining a fixed set of objects helps reduce the variability between experiments carried out on different platforms, dedicated experimental procedures were needed to standardise how to benchmark robotic solutions. For this reason, benchmark protocols have been proposed to evaluate specific components of robotic systems, such as hardware [17], [18], [19], grasp planning algorithms [11], [10], object caging strategies [20] or motion planners [21]. Other benchmarking protocols have been proposed to evaluate the performance of robotic systems as a whole on specific use cases, such as grocery picking [22], cloth folding [23], aerial manipulation [24], shelf picking [25], and others. In addition to a detailed explanation of how to carry out the experiments, all these works introduce task-specific metrics that depend on a specific definition of success, usually related to their use case.

Instead of labelling experiments with task-specific binary outcomes (i.e. success or failure), Bekiroglu et al. [10] have recently proposed a grasp planning benchmark protocol in which several binary labels are assigned to each trial. Each binary label corresponds to whether a generated grasp configuration passes a given sub-test. This approach is able to extract more detailed information (e.g. resiliency to lifting and rotational motions) about the robot's behaviour than a simple success/failure ratio.

Similarly, in order to decouple the limitations of grasp planning algorithms from those of the testing platform, Bottarel et al. recently proposed GRASPA [11]. This benchmarking protocol includes separate assessments of different parts of a given grasping pipeline via a set of sub-metrics (e.g. hardware reachability, camera calibration, graspability, grasp quality and grasp stability). These scores are then combined into an overall metric accounting for the limitations of the robotic platform used to carry out the grasping experiments.

Although these two works describe thorough and reproducible experimental procedures, the proposed metrics and outcomes used to assess the performance of grasp planning algorithms rely on specific definitions of a successful grasp, which are not universal [9]. In addition, none of these works provides any method to statistically rank grasp planning algorithms based on the proposed metrics. For instance, Bekiroglu et al. [10] report the results of two grasping methods but do not provide any statistical argument that one is significantly better than the other. In fact, the conclusions drawn from dichotomous outcomes (simply counting the number of successes) can be misleading, as they do not account for the fact that the same experiments can lead to different outcomes due to some random variability [26]. This makes it difficult to conclude that equivalent performance has really been achieved, by two or more approaches.

For these reasons, we identified the need for a novel approach to compute the statistical ranking of grasp planning algorithms, which is decoupled from the choice of experimental protocols. We believe that to be generic, the ranking of the performance of grasp planning algorithms should be possible regardless of the definition of a successful trial. Hence, we propose in this paper to label the results of an experiment using a stratification of categorical and task-agnostic outcomes that enables the use of well-established statistical frameworks [27]. In addition to allowing researchers to rank the grasp planning algorithms for different definitions of success from a single set of experiments, this statistical method can be used to provide rankings accounting for the effect of multiple experimental variables that can impact the observed performance of grasp planning algorithms. This enables further analysis of the benchmark results that can reveal possible interactions between a grasp planning algorithm and other factors. For example (but not limited to), the proposed statistical framework can be used to determine whether a given grasping method performs better than others, in relation to a particular subset of objects, or for a particular pose of an RGB-D sensor.

## III. Empirical study

In this section, we illustrate the shortcomings of a naive ranking approach, by analysing the performance of four grasp planning algorithms [28], [6], [7], [29]. In particular, we perform the same set of experiments twice, and show that the resulting rankings may vary.

### A. Objects

Different sets of objects have been used in the robotics literature, for the evaluation of both hardware capability and of grasp planning methods. Some objects are more difficult to grasp than others, and this will of course influence the distribution of successful grasps. As previously mentioned, the YCB dataset [12] has been proposed to create a common and reproducible framework to evaluate robotic components for grasping and manipulation. This dataset comprises 72 physical objects for which corresponding high-quality 3D models and RGB-D data are also provided. However, with the advent of data-driven approaches, an increasing number of state-of-the-art methods make use of the provided 3D models and/or RGB-D data to train their grasp planning algorithms [7], [14], [8], [30], [31]. Comparing a set of grasp planning algorithms that make use of the YCB dataset to generate their training data would not necessarily introduce any bias, even though it would not provide meaningful information about their generalisation capacity to unknown objects. On the other hand, if at least one of the methods to be compared does not make use of the YCB resources (e.g. non learning-based approaches), using the YCB physical set of objects would introduce a bias in the evaluation process. Since one of the grasp planning algorithms benchmarked in this work is trained using data from the YCB dataset [7], we

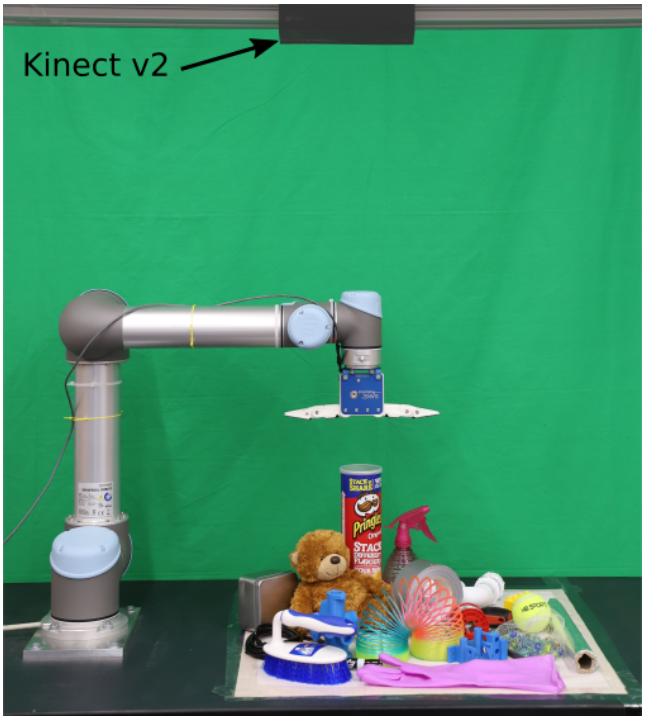



Fig. 2: The robotic setup used to run the protocol described in subsection III-B. All four methods were executed in the same conditions: each of the 20 objects was grasped in isolation, in 5 different poses within a 50×50cm workspace.

have selected a novel set of 20 objects that are *not* part of the YCB set.

As illustrated in Figure 2, this set comprises the following items: a net of marbles, a metallic box, a cardboard tube, a screwdriver, a roll of duct tape, an HDMI cable, a latex cleaning glove, an empty spray bottle, a marker, a socket universal joint (with bars), a tennis ball, a Duplo block, a soft teddy bear, an unpacked roll of kitchen foil, a sink pipe, a brush, a spring toy, a spool of solder, and two 3D printed adversarial objects [32]. We argue that this set of objects exhibits a range of interesting properties such as softness, asymmetry, articulation, deformation, and shininess. Note that we have chosen these objects to be easily purchasable, so that our work can be replicated[1].

### B. Protocol

In order to compare the four grasp planning algorithms, we followed a protocol similar to [10] with an EZGripper mounted on a Universal Robot UR5 arm (see Figure 2). A Kinect2 is located perpendicularly above the workspace, which is defined by a 50cm×50cm square of graph paper. For each repetition, the system starts from a pre-recorded state, from which the robot moves into the defined workspace. The following steps are then performed, on each trial:

- Pre-grasp and grasp pose are generated by the algorithm
- Robot arm moves to the generated pre-grasp pose
- Fingers of the end-effector open to a predefined and constant posture
- Robot arm moves to the generated grasp pose
- Fingers of the end-effector close completely at a constant speed with a limited torque
- Robot moves up to a predefined lifting position and waits for 2 seconds
- Robot executes a predefined and constant trajectory shaking the object (stability test)
- Robot moves back to the generated grasp pose
- Fingers of end-effector open to release the object
- Robot moves back to starting pose

The shaking motion is designed to test the stability of the grasp, bearing in mind that the end-effector is not squeezing with the highest torque possible. Our shaking motion differs from the one described in [10] and is closer to the one described in [11]. In fact, the designed motion reaches successive waypoints within 0.25 seconds, pausing at each one for 0.1 seconds. The trajectory's waypoints are defined around the predefined lifting position. Two waypoints are defined to be 0.175 m above and below this reference pose, while the two others are defined to be 0.16 m to its right and left. More particularly, the shaking motion comprises three up-down motions of an amplitude of 0.35 m each and three left-to-right motions of an amplitude of 0.32 m. This leads to an energetic shake, empirically testing the stability of a generated grasp configuration under dynamical movements of the robot arm. If the object is not grasped, or if it fell from the robot's end-effector before the shaking motion, then the experiment is stopped, and the stability test is not executed.

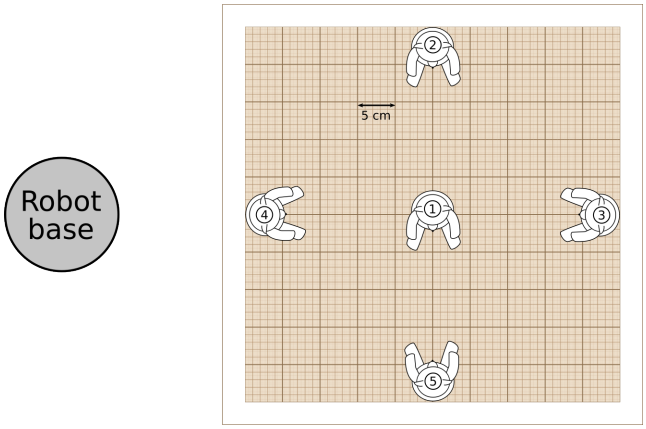



Fig. 3: Schematic view of the five constant poses defined within our 50cm×50cm workspace from which each of the 20 isolated objects will be grasped.

For each of the five poses in which each isolated object is placed (see Figure 3), we repeat this experimental procedure five times, leading to a total of 25 grasps executed per object. Note that this number of repetitions reflects the conditions in which other works evaluate novel grasp planning algorithms [33], [7], [6], where methods are executed between 10 and 15 times on isolated objects. All experiments were carried out using the GRIP framework [34].

[1] www.eecs.qmul.ac.uk/~bdd30/benchmark_presentation

### C. Grasp scoring

As previously mentioned, the definition of a successful grasp depends on the targeted task. For this reason, we design a more informative and generic scoring system that separately evaluates different features of an executed grasp configuration. In fact, only counting the number of successes does not provide any information about how the grasp fails, which can entail crucial information that can be used to determine future research directions. In this work, we consider that attempting a grasp will lead to one of six categorical outcomes, as follows:

- M — Grasp **misses**; object not contacted.
- MC — Grasp **misses**; but object **contacted**.
- U — Grasp **unstable**; object falls when lifted or within two seconds of reaching the stable position.
- DU — Grasp **dynamically unstable**; object falls during the shaking motion, or before being released.
- PS — Grasp is **partially stable**; object deposited on the table, in an arbitrary position.
- S — Grasp is **stable**; object deposited on the table, in its original position.

In more detail, we define a grasp as **stable** if and only if at least half of the object is within its original footprint on the table, and within an angular tolerance of $\pm 45°$ after the robot arm moves back to its starting pose and the gripper is opened. Although the robotics community has proposed different definitions of grasp stability [35], [36], ours relates to whether the pose of the object with respect to the gripper changes significantly during the lifting and shaking of the object. In fact, these outcomes have been designed to measure the underlying time of contact between the gripper and the object during the grasping attempt, which we believe to be an acceptable evaluation of grasp stability. Note that the proposed scoring system does not account for success, which is task-specific [37], but simply describes the result of an experiment that follows this protocol. We argue that using a scoring system less task-oriented makes the reported results more meaningful for a wider number of researchers. In fact, using this rule to describe an outcome still enables researchers to straightforwardly compute the success rate of a given grasp planning algorithm, depending on their use case. It is important to note that our scoring approach is similar to the one adopted in [10], i.e. we evaluate the quality of a grasp at different steps of the experimental procedure instead of only considering its final outcome.

### D. First set of experiments

In this subsection, we are interested in empirically comparing the performance of four grasp planning algorithms [6], [7], [28], [29]. To do so, we follow the protocol described above, leading to a total of 25 grasp attempts per object. Therefore, each grasp planning algorithm is evaluated on $25 \times 20 = 500$ grasps. The resulting distribution of outcomes for each grasping method is depicted in Figure 4.

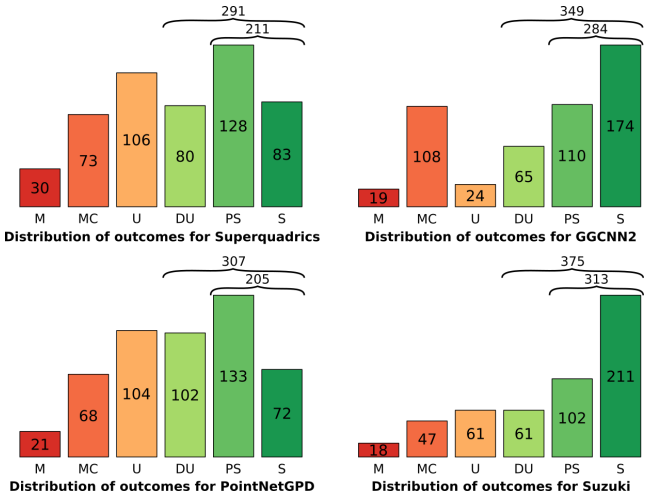



Fig. 4: Results obtained for the first set of experiments. The (leftward) cumulative distribution of the three highest ranking outcomes is represented above the distribution of outcomes of each method considered in this work.

We observe that the outcomes of the 500 generated grasp configurations do not follow the same distribution for each grasp planning algorithm. This suggests that the benchmarked grasp planning algorithms have different characteristics. For instance, GGCNN2 [6] seems to generate more grasp configurations that contact objects but fail to robustly grasp them, compared to other methods. Similarly, PointNetGPD [7] seems to generate a more important proportion of grasp configurations that are not resilient to our shaking motion than any other method.

Interestingly, scoring the grasp outcomes following the scale introduced in section III-C, allows for comparing the success rate of the methods for different definitions of success. If we consider that a generated grasp is successful if the object is grasped and does not fall from the end-effector after 2 seconds, then we can directly compare the total count of outcomes labelled as DU, PS and S. In this case, empirically comparing these numbers leads us to conclude that Suzuki [29] performs the best, followed by GGCNN2 [6], then by PointNetGPD [7], while Superquadrics [28] is the worst performing method. On the other hand, considering success to only be grasps labelled as **stable** leads to a different ranking, as illustrated in Figure 4. In fact, with this stricter definition of success, the Superquadrics method performs better than PointNetGPD, while the two other methods maintain their original ranks. Therefore, it seems that the resulting performance rankings depend on which subset of outcomes of the proposed scoring scale is considered to be a 'success'. This shows the importance of labelling experiments with more than two outcomes.

### E. Second set of experiments

We now replicate the previous experiments in order to establish whether or not the previous rankings of the grasp

planning algorithms are reliable. By following the same protocol, we might expect to observe the same rankings as before. The outcomes observed for the new set of 2000 experiments (i.e. 500 experiments per grasping algorithm) are summarised in Figure 5.

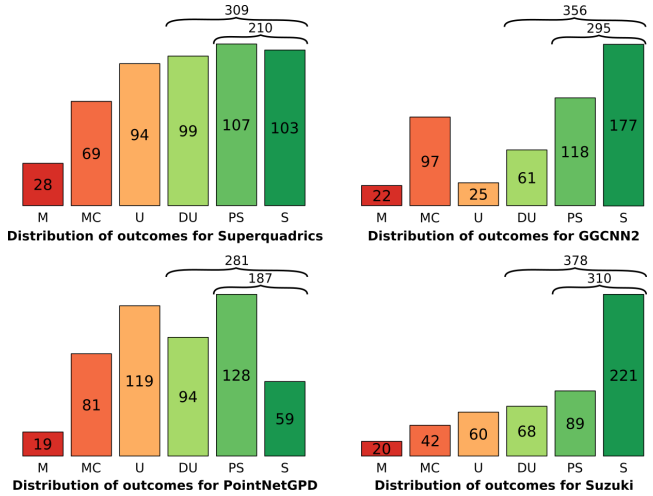



Fig. 5: Results obtained for the second set of experiments. The (leftward) cumulative distribution of the three highest ranking outcomes is represented above the distribution of outcomes of each method considered in this work.

Differences in the outcome distributions of the two sets of experiments can be observed. For instance, the Superquadrics method generated a higher proportion of partially stable (PS) grasp configurations in the second set of experiments. This means that despite all efforts to reduce the variability in the experimental protocol as much as possible, some randomness can be observed in the experiments. For the strictest definition of success (i.e. only S), it appears that such changes in the distributions do not affect the resulting ranking of the grasp planning algorithms. In fact, the empirical ranking is the same as in the previous subsection; Suzuki outperforms all the other methods, followed by GGCNN2, and the superquadrics-based approach is more performant than PointNetGPD. On the other hand, when considering a success to be outcomes labelled as DU, PS and S, the empirical ranking obtained on this set of experiments is different from the one obtained in the previous subsection. In fact, although the rankings of GGCNN2 and Suzuki remain the same, we can empirically conclude that Superquadrics is ranked before PointNetGPD for this set of experiments. This discrepancy shows that empirically comparing the observed numbers can lead to misleading conclusions that are only true for a specific subset of experiments.

As outlined in this study, simply comparing the number of occurrences of specific outcomes is not robust to the inherent variability of the experiments, even when following a dedicated protocol. While one solution to this problem would be to increase the number of repetitions, it would make the benchmark of each grasp planning algorithm much longer, thus drastically increasing the workload related to benchmarking. A more practical solution is to analyse the results using statistical methods that account for such variability. For this reason, we present in the following sections a statistical model that can not only rank grasp planning algorithms, but also account for the effect of other factors (e.g. different target objects) on the observed performance.

## IV. Stratification

As previously mentioned, one of the most common strategies employed to score the outcome of a grasp is to use a dichotomous variable, meaning success or failure [8], [30], [31]. In order to be less biased by the definition of success [9] and to be able to extract more detailed information from the same set of experiments [10], we have proposed in section III-C to score each grasp via a stratification of six qualitative outcomes. In this section, we define how to create a stratification compatible with the statistical framework presented later in this work.

### A. Ordering of outcomes

Although we present in section III-C a stratification to describe the performance of grasp planning algorithms, we argue that researchers should be able to modify it or even create their own set of outcomes. For this reason, we propose an approach to defining these outcomes, while enabling the use of well-established statistical frameworks. To ensure compatibility with the different statistical models mentioned later in this work, the stratification must be composed of outcomes that are mutually exclusive and ordered. In other words, we can define a stratification as

$$O = \{o_1, \dots, o_c\} \text{ where } o_1 \prec \cdots \prec o_c \text{, and } c \geq 2, \quad (1)$$

where $c$ is the number of outcomes (at least 2), and where $o_i \prec o_j$ means that the rank of outcome $o_i$ is lower than that of outcome $o_j$. The result of a given experiment must be associated with a unique outcome from the above sequence. Note that the outcomes must be ordered from the *worst* to the *best*. The order should have some underlying meaning with respect to the experiment or what is being evaluated, e.g. dexterity of a robot hand. For example, the outcomes that were proposed in subsection III-C form a suitable stratification which satisfies the above criteria. In particular, the outcomes $\{o_{\mathsf{M}}, o_{\mathsf{MC}}, o_{\mathsf{U}}, o_{\mathsf{DU}}, o_{\mathsf{PS}}, o_{\mathsf{S}}\}$ are mutually exclusive and ordered according to the amount of time that the end-effector was in contact with the target object. For instance, outcome $o_{\mathsf{M}}$ (i.e. miss) corresponds to no contact between the gripper and the target object while outcome $o_{\mathsf{U}}$ (i.e. object falls before the shaking motion) is considered *worse* than $o_{\mathsf{DU}}$ (i.e. object falls during the shaking motion) since the object is grasped for a shorter amount of time.

Although defining a large number of outcomes would allow researchers to extract more information about the behaviour of a robot, it also increases the difficulty and/or time to label experiments. We argue that the number and meaning of the outcomes should depend on the protocol under which experiments are performed and how fine-grained the analysis aims to be. Note that the set of outcomes defining

the stratification can be designed to evaluate the performance and design of other robotic components. In practice, we believe that a stratification containing at least five outcomes enables researchers to describe a grasp planning algorithm in detail.

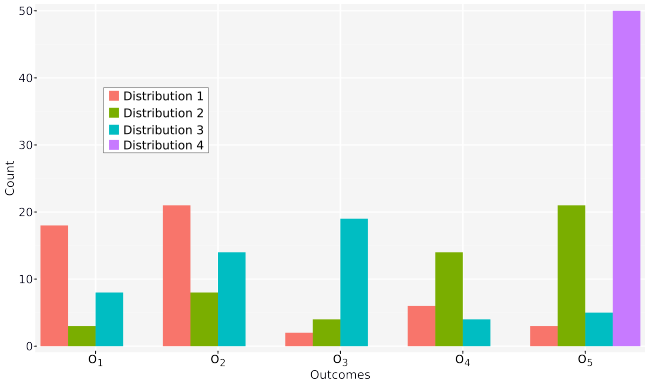
### B. Probability of outcomes

Let $X$ be a categorical variable that can take as value any of the grasp planning algorithms we want to compare; i.e $X \in \{x_1, \dots, x_r\}$, $r \geq 2$. Let $Y$ be a variable which can take as value any outcome of the stratification defined in Equation 1. Let us also define the notation $Y \preceq o_j$, meaning that $Y$ can take any outcome with a rank lower or equal to that of $o_j$.

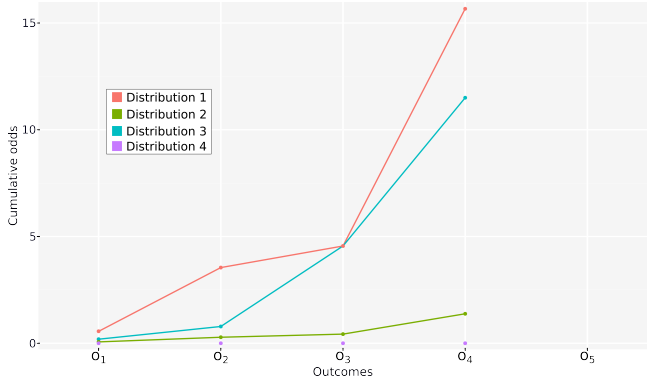
The first purpose of this work is to statistically rank any number of grasp planning algorithms, given a set of observed outcomes. To do so, let us define the cumulative odds to observe an outcome $o_j$ for a method $x_i$, as follows:

$$\theta_i^{o_j} = \frac{P(Y \preceq o_j \,|\, x_i)}{1 - P(Y \preceq o_j \,|\, x_i)} = \frac{P(Y \preceq o_j \,|\, x_i)}{P(Y \succ o_j \,|\, x_i)}, \qquad (2)$$

where $1 \leq j \leq c-1$. The smaller a given $\theta_i^{o_j}$ is, the higher the chances are of observing outcomes with ranks strictly higher than $j$.




(a) Example of 4 distributions of outcomes over a total of 50 trials each.




(b) Cumulative odds of each outcome for the same distributions.

Fig. 6: Illustration of four distributions of outcomes and their corresponding cumulative odds. The smaller the cumulative odds of a given outcome, the higher the probability of observing an outcome with a higher rank.

As illustrated in Figure 6, a set of cumulative odds smaller than 1 for low-ranking outcomes describes a distribution more shifted towards better outcomes. Note that if a given grasp planning algorithm $x_i$ is performing perfectly, then all of its outcomes correspond to $o_c$, i.e.

$$P(Y \preceq o_{c-1} \,|\, x_i) = 0 \;\Leftrightarrow\; \theta_i^{o_j} = 0, \text{ where } 1 \leq j \leq c-1.$$

The second purpose of this work is to provide more in-depth analysis of the benchmarking results, and more particularly to provide rankings that account for the effect of other experimental factors. For instance, if grasp planning algorithms have been evaluated on multiple platforms with different RGB-D cameras, we could define $Z \in \{z_1, \dots, z_b\}$, $b \geq 2$ as the vision sensor used to collect the input data. In this particular example, researchers can be interested in determining if a grasp planning algorithm $x_i$ performs better than the others when provided with inputs captured with a specific camera $z_k$. In this case, we can define the cumulative odds $\theta_{ik}^{o_j}$ of any outcome $o_j$ for a specific combination of variables $x_i$ and $z_k$ from $P(Y \preceq o_j | x_i, z_k)$. Note that cumulative odds can be defined for any number of variables (categorical or continuous). For notational simplicity, it is implicit that $1 \leq j \leq c-1$, whenever the relation $Y \preceq o_j$ appears.

## V. Statistical Analysis

Since each grasp attempt is described by a unique outcome, the results can be reported in contingency tables. Table I shows an example, in which $r \geq 2$ is the number of grasp planning algorithms to be compared, $n_{i,j}$ represents the number of experiments observed to result in outcome $o_j$ for the grasping algorithm $x_i$.

| | Stratification | | | |
|---|---|---|---|---|
| | $o_1$ | $\cdots$ | $o_c$ | Total |
| $x_1$ | $n_{1,1}$ | $\cdots$ | $n_{1,c}$ | $n_{1,*}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_r$ | $n_{r,1}$ | $\cdots$ | $n_{r,c}$ | $n_{r,*}$ |
| Total | $n_{*,1}$ | $\cdots$ | $n_{*,c}$ | $T$ |

TABLE I: Format of a contingency table that reports the observed outcomes for each grasp planning algorithm $x_i$.

We define $n_{*,j}$ as the total number of experiments labelled as $o_j$ across the grasp planning algorithms to be compared. Similarly, $n_{i,*}$ is the total number of experiments carried out for each grasping algorithm. Finally, $T$ is the total number of grasps executed across all $x_i$. In this scenario, $n_{i,*}$ should be equal to the total number of experiments carried out using $x_i$. Note that when considering studies with more than one experimental factor, the results can be described by several contingency tables, each one gathering the number of outcomes observed for a given combination of the other variables. For instance, if we consider a second experimental variable $Z \in \{z_1, \dots, z_b\}$, $b \geq 2$, results can be described in $b$ contingency tables, one per value of $Z$.

### A. Statistical tests

Given a contingency table in the format of Table I, the standard $\chi^2$ test of homogeneity can be applied, based on the statistic

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{i,j}-e_{i,j})^2}{e_{i,j}}, \text{ where } e_{i,j} = \frac{n_{i,*}\times n_{*,j}}{T}.$$

In our case, the null hypothesis $H_0$ of this test is that all grasp planning algorithms lead to the same distribution of outcomes. If the null hypothesis is rejected, then at least one grasping algorithm $x_i$ has a different distribution.

When $r = c = 2$, it is possible to directly draw a conclusion regarding which grasp planning algorithm is *better* than the other. However, when $r \geq 3$ and $c = 2$, identifying which one is *better* requires us to run several pairwise comparisons with appropriate corrections to account for multiple tests. In case $H_0$ is rejected, ranking the grasping algorithms needs a total of $\frac{r\times(r-1)}{2}+1$ tests for each contingency table. Finally, when $c \geq 3$ it is important to note that the analysis becomes more difficult since the test only informs us that there is a difference in the distributions, and not that one is more distributed toward the *best* outcomes (because the $\chi^2$ value does not account for the ordering).

For this reason, multiple tests for ordinal variables have been introduced [38], but are not well suited to our constraints (e.g. assumptions on the normality of distributions). For instance, the Kruskal-Wallis statistic [39] tests if the rank of the median outcome of 2 or more distributions is equivalent. Although this test accounts for the ordered nature of the outcomes, it does not provide a detailed ranking of all the methods, except if coupled with pairwise tests.

### B. Ordinal regression

A more general approach is to regress the observed outcomes on a set of independent variables (e.g. grasp planning algorithms, RGB-D sensors, pose of the objects, etc.). Among the different regression models proposed in the literature [40], the cumulative logit model [41] is the most appropriate. This choice allows us to model the observed distribution of ordinal outcomes as a function of a set of continuous, categorical or ordinal variables. In particular, this regression allows us to model the observed distributions of outcomes $Y$ as a function of each grasp planning algorithms $x_i$ via the following equation:

$$\text{logit}[P(Y \preceq o_j \,|\, x_i)] = \beta_j + \tau_{ij}, \quad 1 \leq i \leq r-1. \tag{3}$$

Note that the above model is directly related to the cumulative odds $\theta_i^{o_j}$ via

$$\text{logit}[P(Y \preceq o_j \,|\, x_i)] = \log\frac{P(Y \preceq o_j \,|\, x_i)}{1-P(Y \preceq o_j \,|\, x_i)} = \log(\theta_i^{o_j}).$$

Using this formulation, each $\beta_j$ (also referred to as intercept) can be defined as $\log(\theta_r^{o_j})$ and corresponds to the logarithm of the cumulative odds of observing $o_j$ for the grasp planning algorithm $x_r$. Similarly, each coefficient $\tau_{ij}$ represents the effect of grasp planning $x_i$ on the logarithm of the cumulative odds of observing $o_j$.

Although modelling the effect of grasp planning algorithms independently for each outcome allows us to carry out a detailed analysis, this approach cannot be used with additional factors [27]. In particular, this model is reported to have the structural problem [27] that cumulative probabilities can be out of order when modelling the joint effect of additional factors.

For this reason, a more constrained model, often referred to as ordered logit (or ordinal logistic) regression, has been proposed. It assumes the proportional odds property, i.e. the effect of each factor is identical for each of the $c-1$ cumulative probabilities. Although this model tends to fit the data less well than model (3), especially when the proportional odds assumption is violated, it is known that this simpler model remains valid for comparing the overall tendencies of ordinal variables [27]. In other words, such models allow for ranking the different levels of the variable $X$ based on the overall tendency of the observed outcomes of a given stratification. If we consider ranking the distribution of observed outcomes according to both grasping algorithms and another variable $Z \in \{z_1, \ldots, z_b\}$, $b \geq 2$, then the model is

$$\text{logit}[P(Y \preceq o_j \,|\, x_i, z_k)] = \beta_j + \tau_i + \eta_k + \phi_{ik}, \tag{4}$$

for $i = 1, \ldots, r-1$ and $k = 1, \ldots, b-1$. In this model, each intercept corresponds to the logarithm of the cumulative odds of observing $o_j$ for grasp planning $x_r$ when $Z = z_b$. Here, $\tau_i$ corresponds to the effect of the grasping method $x_i$ over the logarithm of any of the $c-1$ cumulative odds when estimated for $Z = z_b$. Similarly, $\eta_k$ corresponds to the effect of $z_k$, for the grasping algorithm $x_r$, on any of the $c-1$ cumulative log odds. Note that Equation 4 contains a fourth term, $\phi_{ik}$ that models any possible effect of the interaction between the grasp planning algorithm $x_i$ and $z_k$ on the $c-1$ cumulative log odds.

Note that ordinal logistic regression can also describe the interaction between more than two factors. For instance, to account for the impact of two additional experimental variables ($Z$ and $W \in \{w_1, \ldots, w_d\}$, $d \geq 2$) on the observed distribution of outcomes for $r$ grasp planning algorithms, the model becomes

$$\log(\theta_{ikl}^{o_j}) = \beta_j + \tau_i + \eta_k + \gamma_l + \phi_{ik} + \lambda_{il} + \mu_{kl} + \psi_{ikl}, \tag{5}$$

for $i = 1, \ldots, r-1$, $k = 1, \ldots, b-1$ and $l = 1, \ldots, d-1$. In this model, we can define $\beta_j$ as $\log(\theta_{rbd})$, i.e. as the logarithm of the cumulative odds of observing each outcome $o_j$. Following the same logic as previously, $\tau_i, \eta_k$ and $\gamma_l$ represent the effect of the corresponding variable when the two other variables are set to their respective reference values. For instance, $\tau_i$ corresponds to the effect of grasp planning algorithm $x_i$ when $Z = z_b$ and $W = w_d$. The terms $\phi_{ik}, \lambda_{il}$ and $\mu_{kl}$ model the pairwise interaction effect between two variables when the third is set to its reference value. For instance, $\mu_{kl}$ represents the interaction between $z_k$ and $w_l$ for $X = x_r$. Finally, the last term $\psi_{ikl}$ represents the effect of any potential three-way interaction between $x_i$, $z_k$

and $w_l$ on the logarithm of the cumulative odds of observing any $o_j$, $1 < j < c-1$.

### C. Coefficient interpretations

Model (3) outputs a set of $c-1$ intercepts $\beta_j$ and $(c-1)(r-1)$ coefficients $\hat{\tau}_{ij}$. Given the definition of $\beta_j$, the coefficients $\hat{\tau}_{ij}$ can be expressed as

$$\hat{\tau}_{ij} = \log(\theta_i^{o_j}) - \log(\theta_r^{o_j}) = \log\left(\frac{\theta_i^{o_j}}{\theta_r^{o_j}}\right), \; i = 1, \dots, r-1.$$

Hence the exponentials of these coefficients can be used to compare the cumulative odds of any outcome $o_j$ for each grasping algorithm $x_i$ to that of $x_r$. If $\exp(\hat{\tau}_{ij}) > 1$ is statistically significant, we can conclude that the grasp planning algorithm $x_r$ leads to more outcomes with a rank strictly higher than $j$ than the method $x_i$. Note that a single regression contains the information required to statistically rank all of the grasping algorithms, according to their distribution of outcomes. In fact, for $\hat{\tau}_{aj}$ and $\hat{\tau}_{bj}$, we have

$$\hat{\tau}_{aj} - \hat{\tau}_{bj} = \log\left(\frac{\theta_a^{o_j}}{\theta_r^{o_j}}\right) - \log\left(\frac{\theta_b^{o_j}}{\theta_r^{o_j}}\right) = \log\left(\frac{\theta_a^{o_j}}{\theta_b^{o_j}}\right).$$

If the observed number of outcomes with a rank smaller or equal to the one of $o_j$ is statistically similar for method $a$ and for method $b$, then the value

$$z^2 = \frac{(\hat{\tau}_{aj} - \hat{\tau}_{bj})^2}{\mathrm{Var}(\hat{\tau}_{aj} - \hat{\tau}_{bj})}$$

follows a $\chi_1^2$ distribution. Therefore, the output of a single regression allows for testing whether the difference of effect between any pair of grasp planning algorithms is statistically significant, for the cumulative odds of observing each outcome $o_j$.

For model (4), due to the proportional odds assumption, each $\beta_j$ corresponds to the logarithm of cumulative odds of each outcome $o_j$ when $X = x_r$ and $Z = z_b$. In this scenario, the exponential of $\hat{\tau}_i$ enables us to directly compare the distribution of outcomes of the grasp planning algorithm $x_i$ to $x_r$ for $Z = z_b$. Similarly, the exponential of each $\hat{\eta}_k$ allows us to compare the distribution of observed outcomes between $z_k$ and $z_b$ when $X = x_r$. On the other hand, $\exp(\hat{\phi}_{ik})$ being statistically different than 1 means that there is an interaction between $z_k$ and $x_i$, and does not directly inform us whether the grasp planning algorithm $x_i$ performs better than $x_r$ for $Z = z_k$. To perform such a comparison, we need to compute the logarithm of the cumulative odds ratio of $\theta_{ik}^{o_j}$ and $\theta_{rk}^{o_j}$ from Equation 4

$$\log\left(\frac{\theta_{ik}^{o_j}}{\theta_{rk}^{o_j}}\right) = \hat{\beta}_j + \hat{\tau}_i + \hat{\eta}_k + \hat{\phi}_{ik} - (\hat{\beta}_j + \hat{\eta}_k) = \hat{\tau}_i + \hat{\phi}_{ik}.$$

Therefore, if $\exp(\hat{\tau}_i + \hat{\phi}_{ik}) < 1$ is statistically significant, then we can conclude that algorithm $x_i$ will generate overall better outcomes than $x_r$ when $Z = z_k$. Similarly, we can determine which grasp planning algorithm $x_i$ or $x_{i'}$ performs better for a given $z_k$ by determining if $\exp(\hat{\tau}_i + \hat{\phi}_{ik} - \hat{\tau}_{i'} - \hat{\phi}_{i'k})$ is statistically different than 1. We argue that the coefficients output by the model can also provide further insight regarding the performance of each grasp planning algorithm in isolation. In fact, if researchers are interested in understanding whether a given grasp planning algorithm $x_i$ performs similarly in condition $z_k$ and $z_{k'}$, the statistical conclusion can be obtained by testing the statistical significance of any departure from $\exp(\hat{\eta}_k + \hat{\phi}_{ik} - \hat{\eta}_{k'} - \hat{\phi}_{ik'}) = 1$. If such a study needs to be carried out for $X = x_r$, the values to be computed are simplified to $\exp(\hat{\eta}_k - \hat{\eta}_{k'})$.

Naturally, the number of analyses accounting for affinities that are enabled by the ordinal logistic regression increases with the number of experimental variables considered. For instance, the set of $\hat{\tau}_i + \hat{\phi}_{ik} + \hat{\lambda}_{il} + \hat{\psi}_{ikl}$, $\hat{\tau}_i + \hat{\lambda}_{il}$ and $\hat{\tau}_i + \hat{\phi}_{ik}$ can be computed from the output of model (5) and allows researchers to rank the performance of the grasping algorithms for any value of $Z$ and $W$. An example of such a study is reported in section VI-D. In order to facilitate the use of this statistical framework, we provide an R script that computes all the aforementioned coefficients and p-values to either rank grasping algorithms or study the interaction between multiple experimental variables from a single regression[2].

## VI. Case studies

In this section, we demonstrate the benefits of using ordinal regression to analyse the results of benchmarks. First, we show in section VI-A that unlike the empirical approach, rankings of grasp planning algorithms using a statistical model are more robust to small numbers of repetitions. Second, we demonstrate three possible in-depth analyses of the observed distributions of outcomes based on the same set of experiments. In section VI-B, we extract object-specific rankings of the performance of grasp planning algorithms. In section VI-C, we present a complementary analysis based on the outputs of the same ordinal regression that enables us to study the affinity of a given grasp planning algorithm to the set of test objects. Last, we show in section VI-D an example of rankings obtained when accounting for a third experimental variable; the pose of the grasped objects.

### A. Comparison of grasp planning algorithms performance

As mentioned in section III-E, the empirical rankings obtained over two sets of experiments run under the same conditions do not match. One of the underlying reasons might be the relatively low number of repetitions given the different conditions (i.e. poses and objects). To demonstrate the robustness of the proposed approach to relatively small numbers of repetitions, a third set of experiments is carried out following the exact same protocol as the two presented in section III, and the rankings obtained by the two methods are compared across the three sets of experiments. The observed distributions of outcomes are reported in Figure 7, which is a graphical representation of the corresponding contingency table.

[2] https://github.com/ARQ-CRISP/stratified_statistical_analysis.git

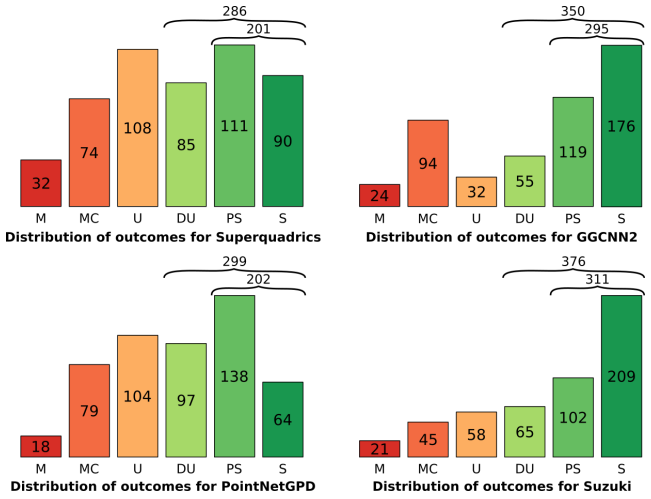



Fig. 7: Results obtained for the third set of experiments. The cumulative distribution of the three highest ranking outcomes is represented above the distribution of outcomes of each method considered in this work.

Since this study aims at ranking the performance of grasp planning algorithms over the entire set of objects (i.e. regardless of which object is being considered), we fit model (3) to the collected outcomes. The output coefficients of the regression are reported in Table II.

| Effect | $o_{\mathsf{M}}$ | $o_{\mathsf{MC}}$ | $o_{\mathsf{U}}$ | $o_{\mathsf{DU}}$ | $o_{\mathsf{PS}}$ |
|---|---|---|---|---|---|
| $\hat{\tau}_1$ | 0.442 | **0.568** | **0.817** | **0.892** | **1.180** |
| $\hat{\tau}_2$ | 0.138 | **0.706** | 0.259 | 0.131 | **0.274** |
| $\hat{\tau}_3$ | −0.162 | **0.457** | **0.709** | **0.884** | **1.582** |
| $\hat{\tau}_1 - \hat{\tau}_2$ | 0.305 | −0.138 | **0.557** | **0.761** | **0.906** |
| $\hat{\tau}_1 - \hat{\tau}_3$ | **0.605** | 0.111 | 0.107 | 0.008 | **−0.402** |
| $\hat{\tau}_2 - \hat{\tau}_3$ | 0.300 | 0.249 | **−0.450** | **−0.753** | **−1.308** |

TABLE II: Parameters estimated by model (3) for the third set of experiments. Significant values for $\alpha = 0.05$ are reported in bold.

The effect of the Suzuki grasp planning algorithm has been set as the reference. The coefficients can be interpreted as follows. The cumulative odds of observing outcome $o_{\mathsf{MC}}$ are $\exp(0.706) = 2.03$ times greater for GGCNN2 than for Suzuki, which means that the latter statistically generates more grasp configurations labelled as $\{\mathsf{U}, \mathsf{DU}, \mathsf{PS}, \mathsf{S}\}$ than the former. On the other hand, although the cumulative odds of observing $o_{\mathsf{U}}$ for Superquadrics is $\exp(0.107) = 1.11$ times the odds of PointNetGPD, we cannot statistically conclude that the number of generated grasps $\in \{\mathsf{DU}, \mathsf{PS}, \mathsf{S}\}$ is different for these two grasp planning algorithms.

Note that interpreting the last three columns of Table II allows us to compare the four grasp planning algorithms, for three distinct definitions of a successful grasp, from only one set of experiments and one regression. For instance, if any outcome $\{\mathsf{DU}, \mathsf{PS}, \mathsf{S}\}$ is considered a success, then looking at column $o_{\mathsf{U}}$ enables us to draw a conclusion regarding the ranking of the grasp planning algorithms. In this case, we can observe that Suzuki and GGCNN2 both perform statistically better than Superquadrics and PointNetGPD. However, our approach does not find any statistical difference between the performance of Superquadrics and PointNetGPD or between Suzuki and GGCNN2. In other words, our model attributes the difference in the total number of outcomes with a rank strictly higher than $o_{\mathsf{U}}$ between Superquadrics and PointNetGPD to the observed randomness across the repetitions of the grasp attempts. Interestingly, the empirical rankings observed using the data of the two first sets of experiments seem to confirm this, as we obtained two different rankings for the same definition of success. In order to highlight the benefits of the proposed ordinal logistic regression over the empirical strategy, we can compare the rankings obtained by the two approaches for each set of experiments in Table III.

| | Superquadrics | | | | | |
|---|---|---|---|---|---|---|
| | Set #1 | | Set #2 | | Set #3 | |
| | Emp. | Stat. | Emp. | Stat. | Emp. | Stat. |
| $\{\mathsf{DU}, \mathsf{PS}, \mathsf{S}\}$ | 4 | 3 | 3 | 3 | 4 | 3 |
| $\{\mathsf{PS}, \mathsf{S}\}$ | 3 | 3 | 3 | 3 | 4 | 3 |
| $\{\mathsf{S}\}$ | 3 | 3 | 3 | 3 | 3 | 3 |
| | GGCNN2 | | | | | |
| | Set #1 | | Set #2 | | Set #3 | |
| | Emp. | Stat. | Emp. | Stat. | Emp. | Stat. |
| $\{\mathsf{DU}, \mathsf{PS}, \mathsf{S}\}$ | 2 | 1 | 2 | 1 | 2 | 1 |
| $\{\mathsf{PS}, \mathsf{S}\}$ | 2 | 1 | 2 | 1 | 2 | 1 |
| $\{\mathsf{S}\}$ | 2 | 2 | 2 | 2 | 2 | 2 |
| | PointNetGPD | | | | | |
| | Set #1 | | Set #2 | | Set #3 | |
| | Emp. | Stat. | Emp. | Stat. | Emp. | Stat. |
| $\{\mathsf{DU}, \mathsf{PS}, \mathsf{S}\}$ | 3 | 3 | 4 | 3 | 3 | 3 |
| $\{\mathsf{PS}, \mathsf{S}\}$ | 4 | 3 | 4 | 3 | 3 | 3 |
| $\{\mathsf{S}\}$ | 4 | 4 | 4 | 4 | 4 | 4 |
| | Suzuki | | | | | |
| | Set #1 | | Set #2 | | Set #3 | |
| | Emp. | Stat. | Emp. | Stat. | Emp. | Stat. |
| $\{\mathsf{DU}, \mathsf{PS}, \mathsf{S}\}$ | 1 | 1 | 1 | 1 | 1 | 1 |
| $\{\mathsf{PS}, \mathsf{S}\}$ | 1 | 1 | 1 | 1 | 1 | 1 |
| $\{\mathsf{S}\}$ | 1 | 1 | 1 | 1 | 1 | 1 |

TABLE III: Ranking of each grasp planning algorithm obtained using empirical and statistical analysis for three definitions of success for three sets of experiments consisting of 500 trials each.

The content of Table III shows that rankings obtained via ordinal regression are generally more consistent across repetitions of the same experiments, regardless of the definition of success. This means that the proposed statistical model accounts for some randomness that can be observed during the experiments, and leads to more robust rankings for the same amount of repetitions than the empirical approach. However, the ordinal regression does not seem able to rank GGCNN2 and Suzuki for two definitions of success, while we can empirically observe that Suzuki has consistently a higher number of good outcomes than GGCNN2. This effect can come from two factors; the regression wrongly attributes the small difference to the randomness, or we have not been able to detect a change in the rankings, given only three repetitions of the experiments.

### *B. Object-specific performance ranking of grasp planning algorithms*

Ranking the performance of grasp planning algorithms on a variety of objects allows researchers to draw conclusions about which methods perform overall the best. However, this conclusion does not guarantee that the best-performing grasping algorithm $x_i$ is actually outperforming the others for each object. Determining if that's the case would inform researchers regarding which grasp planning algorithm to select when deployed in environments with specific objects. Although this study can be carried out on a subset of experiments (i.e. 2000 grasps), we make use of the whole set of 6000 grasp executed across the 20 objects[1]. More specifically, we let $Z$ be the variable representing the target object of each experiment. Given the relatively high number of objects considered in this study, model (4) infers a total of 84 coefficients. For clarity, Table IV reports the coefficients enabling us to rank the performance of grasp planning algorithms for a subset of 6 objects.

| | Duct tape | Glove | Marble Net | Metal box | Solder spool | HDMI cable |
|---|---|---|---|---|---|---|
| Superquadrics vs Suzuki | **0.86** | 0.02 | 0.24 | **1.46** | **1.26** | 0.28 |
| GGCNN2 vs Suzuki | **1.36** | **-1.13** | **-0.37** | 0.09 | **0.70** | 0.45 |
| PointNetGPD vs Suzuki | **1.59** | **-1.20** | 0.22 | **1.28** | **1.61** | **0.53** |
| Superquadrics vs GGCNN2 | **-0.50** | **1.15** | **0.61** | **1.37** | **0.56** | -0.17 |
| Superquadrics vs PointNetGPD | **-0.73** | **1.22** | 0.02 | 0.18 | -0.35 | -0.25 |
| GGCNN2 vs PointNetGPD | -0.23 | 0.07 | **-0.59** | **-1.19** | **-0.91** | -0.08 |

TABLE IV: Comparison of the four grasping algorithms performance for six objects, based on the output of model (4). Bold values correspond to significant values for $\alpha = 0.05$.

In this table, the Suzuki method has been used as the reference algorithm (i.e. defined as $x_r$) and *Duct tape* has been set as the reference object (i.e. defining $z_b$). For this reason, the values of the first column correspond to the estimated set of $\hat{\tau}_i$, i.e. $\hat{\tau}_2 = 1.36$ and $\hat{\tau}_1 - \hat{\tau}_3 = -0.73$. To compare the distribution of observed outcomes of each method $x_i$ vs $x_r$ for an object $Z = z_k$, the values of $\tau_i + \phi_{ik}$ were computed. Therefore, comparing two grasp planning algorithms $x_i$ and $x_{i'}$ (with $i$ and $i' \neq r$) for an object $z_k$ (with $k \neq b$), is equivalent to subtracting the values comparing each grasp algorithm to $x_r$ for $z_k$. For example, the value of the coefficient that allows us to compare the distribution of outcomes of the GGCNN2 method and PointNetGPD for the *Metal box* object is $\tau_2 + \hat{\phi}_{24} - (\tau_3 + \hat{\phi}_{34}) = -1.19$. Since this value is statistically significant, we can conclude that GGCNN2 generates overall grasp configurations that would rank higher than the one generated by PointNetGPD for this particular object. By interpreting each column of Table IV, the rank of the four grasp planning algorithms for each object can be determined, as reported in Table V.

| | Superquadrics | GGCNN2 | PointNetGPD | Suzuki |
|---|---|---|---|---|
| Duct tape | 2 | 3 | 3 | 1 |
| Glove | 3 | 1 | 1 | 3 |
| Marble net | 2 | 1 | 2 | 2 |
| Metal Box | 3 | 1 | 3 | 1 |
| Solder spool | 3 | 2 | 3 | 1 |
| HDMI cable | 1 | 1 | 1 | 1 |

TABLE V: Rankings (best to worst) of the four grasping algorithms for a subset of six objects. The occurrence of tied ranks (in a given row) means that the corresponding differences are statistically insignificant.

Although Suzuki and GGCNN2 were mostly ranked first in the previous study, we can observe that they don't maintain this rank for all of the reported objects. For instance, when considering *Metal box*, GGCNN2 and Suzuki appears to equally perform the best, followed by PointNetGPD and Superquadrics at equal rank. On the other hand, for *Solder spool*, Suzuki is ranked first, followed by GGCNN2, then both Superquadrics and PointNetGPD. Interestingly, for *HDMI cable*, the logistic regression does not allow a clear ranking between the evaluated algorithms, and therefore should be interpreted as all the methods performing on a similar level for this specific object. Although this model can not be interpreted for a specific definition of success, it allows researchers to compare which algorithms generate overall better-ranked grasp configurations for each object. However, relevant graphical representations such as ternary plots (see Figure 8) can give a better overview of the affinity of grasp planning algorithms for specific objects.

Although not extensively explored, such studies could help identify a set of common features from the objects for which two previously equally ranked grasp planning algorithms could be differentiated. For instance, it seems that GGCNN2 is better than Suzuki for flat objects (e.g. glove, marble net) while Suzuki outperforms GGCNN2 on toroidal-like objects (e.g. solder spool, duct tape).

### *C. Study of the affinity between grasping algorithms and target objects*

Although the previous study tells us if a grasp planning algorithm $x_i$ shows a better performance than the others for a specific object, it is difficult to extract any information regarding whether $x_i$ actually performs well on each object or not. In fact, even though a grasp planning algorithm $x_i$ is ranked first for a given object $z_k$, the number of high-ranking outcomes observed for this object might remain low.

As previously mentioned, the affinity between all grasp planning algorithms and target objects can be concurrently extracted from model (4). More particularly, the affinity of a grasp planning algorithm $x_i \neq x_r$ and all the objects $z_k$ can be ranked from the sets of output $\eta_k + \phi_{ik}$. For $x_i = x_r$, the same information can be extracted from the sets of $\eta_k$. Interpreting these coefficients for each grasp planning algorithm results in a varied number of categories

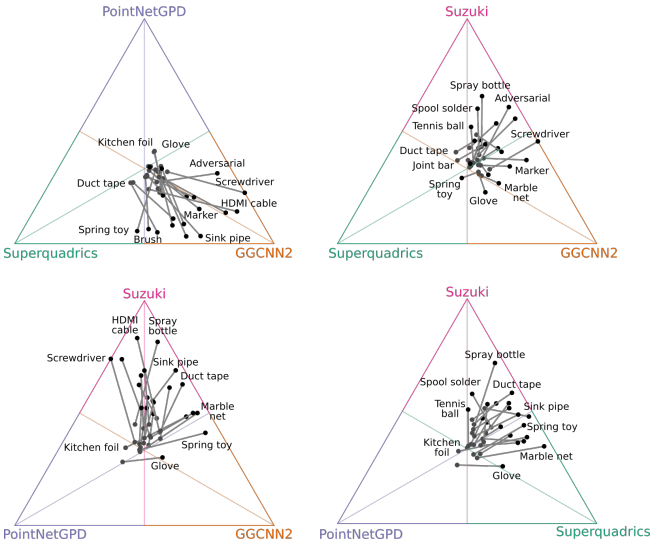



Fig. 8: Ternary plots, showing the proportion of successful grasps executed for each object, with different methods represented by the vertices of the triangles. If the successful grasps of a given object were evenly split between all three methods, then the object-point appears in the centre of the triangle. If the successful grasps were evenly split between two methods, then the object-point appears on the corresponding edge of the triangle. If the successful grasps were dominated by one method, then the object-point appears at the corresponding vertex. The change of object-location, after changing the strictness of a successful grasp, is represented by the link between the grey (low strictness) and black (high strictness) points.

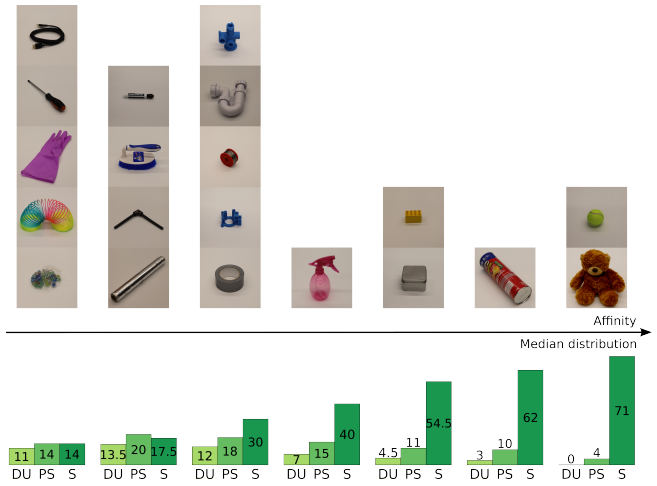



Fig. 9: Graphical representation of the affinity of the Suzuki grasping algorithm to the tested objects. The median distribution of the three highest-ranking outcomes of each *affinity* group inferred by the model is represented. In this figure, the number of grasps executed on each object is 75. We can observe that the median number of high-ranking outcomes increases with the level of affinity inferred by the model.

of objects corresponding to different affinities. A graphical representation of such an affinity scale is illustrated in Figure 9 for the Suzuki method [29].

In the above figure, objects being grouped together is the result of non-significant coefficients. Note that the median distribution of the three highest-ranking outcomes improves with the level of affinity inferred by the proposed statistical analysis. This demonstrates that model (4) enables researchers to study the affinity between objects and grasp planning algorithms in a statistical sense.

We can also observe that both *Glove* and *Marble net*, for which Suzuki ranks 3 and 2 (see Table V) belong to the group of objects for which the grasp planning method has the lowest affinity. On the other hand, *Duct tape*, *Metal box* and *Solder spool* are all part of affinity groups of higher ranks. Interestingly, the object *HDMI cable*, for which the previous study could not rank the performance of grasp planning algorithms, is also part of the group showing the worst affinity. By carrying the same study on all of the grasp planning algorithms, we are able to determine that some objects (i.e. *HDMI cable* and *Screwdriver*) always fall into the lowest-ranked affinity group. This allows us to identify these two objects as the most challenging of the objects used in this specific benchmark. In addition, such information can help highlight possible future works to improve the performance of grasp planning algorithms, e.g. adding instances of geometries similar to the objects identified as challenging in their training dataset.

### D. Rankings accounting for test objects and their pose

In this section, we provide an example of a study that involves two additional explanatory variables that can impact the observed performance of grasp planning algorithms. More complex studies are particularly meaningful to better understand if rankings resulting in many ties are not due to another effect that was not modelled. In our benchmark, another variable that can impact the observed number of high-ranking outcomes is the pose of the object with respect to the camera. For this reason, in addition to accounting for $Z$ (i.e. test object), we also model $W$ as the pose of the object. Here $W$ is a categorical variable, where each $w_l$ corresponds to one of the poses illustrated in Figure 3. Fitting model (5) with the observed outcomes collected on the total set of 6000 grasp experiments results in a total of 404 coefficients. Therefore we report in Table VI the set of $\tau_i + \phi_{ik}$ (for *Pose 1*, set as $w_d$) and $\tau_i + \phi_{ik} + \lambda_{il} + \psi_{ikl}$ (for the other poses) that allows us to rank the performance of the grasp planning algorithms for each pose of two objects; *HDMI cable* and *Marble net*. Note that in the results presented in section VI-B, we established that for the object *Marble net* Suzuki [29], PointNetGPD [7]. and Superquadrics [28] were tied for rank 2. The same study also concluded that all four grasp planning algorithms were performing similarly when tested on *Metal box*.

We can observe that for *Marble Net*, GGCNN2 significantly outperforms Suzuki only when the object is placed in *Pose 2*. Similarly, GGCNN2 outperforms Superquadrics only for *Pose 2* and *Pose 3*, while it performs better than

| | Pose 1 | | Pose 2 | | Pose 3 | | Pose 4 | | Pose 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Marble net | HDMI cable | Marble net | HDMI cable | Marble net | HDMI cable | Marble net | HDMI cable | Marble net | HDMI cable |
| Superquadrics vs Suzuki | 0.47 | 0.20 | 0.21 | -0.50 | 0.75 | **1.94** | 0.02 | 0.56 | -0.15 | 0.48 |
| GGCNN2 vs Suzuki | 0.34 | -0.08 | **-1.48** | -0.47 | -0.56 | **4.93** | 0.11 | **1.39** | -0.49 | **-2.29** |
| PointNetGPD vs Suzuki | 0.88 | -0.21 | -0.61 | 0.16 | **1.73** | **2.71** | **-1.42** | 1.01 | 0.69 | 0.33 |
| Superquadrics vs GGCNN2 | 0.13 | 0.28 | **1.69** | -0.03 | **1.31** | **-2.99** | -0.09 | -0.83 | 0.34 | **2.77** |
| Superquadrics vs PointNetGPD | -0.41 | 0.41 | 0.82 | -0.66 | -0.98 | -0.77 | **1.45** | -0.46 | -0.84 | 0.15 |
| GGCNN2 vs PointNetGPD | -0.54 | 0.13 | -0.87 | -0.63 | **-2.29** | **2.22** | **1.53** | 0.38 | -1.18 | **-2.62** |

TABLE VI: Comparison of the four grasping algorithms performance of two objects accounting for their pose, based on the output of model (5). Bold values correspond to significant values for $\alpha = 0.05$.

PointNetGPD only for *Pose 3*. On the other hand, for both *Pose 1* and *Pose 5*, all the grasp planning algorithms perform similarly on this object. In other words, the ranking obtained for this specific object in the study of section VI-B is mainly due to GGCNN2 performing the best on two poses. Although all grasp planning algorithms seemed to perform similarly on *HDMI cable* in a previous study, the output of model (5) enables us to rank the performance of some grasp planning algorithms for two poses. In particular, Suzuki outperforms all the other grasp planning algorithms only when the object is located on *Pose 3*, while when in *Pose 5*, the best performing grasping method for *HDMI cable* is GGCNN2. For the three other poses, all grasp planning algorithms are reported to perform similarly, which concurs with the rankings obtained in section VI-B.

Therefore, it seems that all of the grasp planning algorithms perform the same when these two specific objects are placed in the centre of the workspace. Although not explored in this work, a more in-depth study similar to the one presented in section VI-C could be carried out to better characterise the affinity between the three variables. We believe that such analysis can help understand if the performance observed for a specific set of conditions $z_k, w_l$ across all algorithms (i.e. resulting in ties in the rankings) can be identified as equally challenging.

## VII. Conclusion

In this work, we have introduced a novel and generic stratified approach to scoring the outcomes of robotic grasp executions, which enables a statistical analysis of the observed results at different levels. Firstly, we demonstrated that ordinal regression enables us to rank the performance of grasping algorithms accounting for the observed randomness across the experiments. Secondly, we demonstrate that the same model can be used to carry out a multi-level analysis of the algorithms that are being compared. Although demonstrated for two specific experimental variables, the studies enabled by the proposed approach can be applied to any other factor impacting the performance of grasp planning algorithms (e.g. RGB-D sensor, end-effector, etc.). We argue that such tools are necessary to better understand the performance of autonomous grasping systems and identify future research directions to improve their performance.

Although not demonstrated in this paper, we believe that our proposed approach is applicable to a wide range of robotics benchmarks. In particular, our statistical procedure can be applied to other stratifications, specifically designed to evaluate different components of robotic systems (e.g. design of end-effector, or choice of motion planning algorithm). Finally, note that we have not inferred anything about the underlying causes of the observed grasp outcomes. For example, it may be that the performance rankings are driven by simple mechanical relations, such as the proximity of the grasp axis to the object centroid. Future work could examine such correlations, in order to understand exactly what makes a successful grasping algorithm.

## References


[1] C. C. Kemp, A. Edsinger, and E. Torres-Jara, "Challenges for robot manipulation in human environments [grand challenges of robotics]," *IEEE Robotics & Automation Magazine*, vol. 14, no. 1, pp. 20–29, 2007.

[2] M. Controzzi, C. Cipriani, and M. C. Carrozza, "Design of artificial hands: A review," in *The Human Hand as an Inspiration for Robot Hand Development*, pp. 219–246, Springer, 2014.

[3] M. Danielczuk, M. Matl, S. Gupta, A. Li, A. Lee, J. Mahler, and K. Goldberg, "Segmenting unknown 3D objects from real depth images using Mask R-CNN trained on synthetic data," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7283–7290, IEEE, 2019.

[4] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," in *Conference on Robot Learning*, pp. 306–316, PMLR, 2018.

[5] 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.

[6] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *The International journal of robotics research*, vol. 39, no. 2-3, pp. 183–201, 2020.

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

[8] Y. Qin, R. Chen, H. Zhu, M. Song, J. Xu, and H. Su, "S4G: Amodal Single-view Single-Shot SE(3) Grasp Detection in Cluttered Scenes," in *Conference on robot learning*, pp. 53–65, PMLR, 2020.

[9] J. Mahler, R. Platt, A. Rodriguez, M. Ciocarlie, A. Dollar, R. Detry, M. A. Roa, H. Yanco, A. Norton, J. Falco, *et al.*, "Guest editorial open discussion of robot grasping benchmarks, protocols, and metrics," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 4, pp. 1440–1442, 2018.

[10] Y. Bekiroglu, N. Marturi, M. A. Roa, K. J. M. Adjigble, T. Pardi, C. Grimm, R. Balasubramanian, K. Hang, and R. Stolkin, "Benchmarking protocol for grasp planning algorithms," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 315–322, 2019.

[11] F. Bottarel, G. Vezzani, U. Pattacini, and L. Natale, "GRASPA 1.0: GRASPA is a robot arm grasping performance benchmark," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 836–843, 2020.

[12] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: The YCB object and model set and benchmarking protocols," *arXiv preprint arXiv:1502.03143*, 2015.

[13] 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. 11444–11453, 2020.

[14] B. Zhao, H. Zhang, X. Lan, H. Wang, Z. Tian, and N. Zheng, "RegNet: region-based grasp network for single-shot grasp detection in point clouds," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[15] K. Matheus and A. M. Dollar, "Benchmarking grasping and manipulation: Properties of the objects of daily living," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5020–5027, IEEE, 2010.

[16] C. Sprunk, J. Röwekämper, G. Parent, L. Spinello, G. D. Tipaldi, W. Burgard, and M. Jalobeanu, "An experimental protocol for benchmarking robotic indoor navigation," in *Experimental Robotics*, pp. 487–504, Springer, 2016.

[17] L. Jamone, A. Bernardino, and J. Santos-Victor, "Benchmarking the grasping capabilities of the iCub hand with the YCB object and model set," *IEEE Robotics & Automation Letters*, vol. 1, no. 1, pp. 288–294, 2016.

[18] F. Negrello, M. Garabini, G. Grioli, N. Tsagarakis, A. Bicchi, and M. Catalano, "Benchmarking resilience of artificial hands," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8374–8380, IEEE, 2019.

[19] I. Llop-Harillo, A. Pérez-González, J. Starke, and T. Asfour, "The Anthropomorphic Hand Assessment Protocol (AHAP)," *Robotics and Autonomous Systems*, vol. 121, 2019.

[20] R. R. Ma, W. G. Bircher, and A. M. Dollar, "Modeling and evaluation of robust whole-hand caging manipulation," *IEEE Transactions on Robotics*, 2019.

[21] M. Moll, I. A. Sucan, and L. E. Kavraki, "Benchmarking motion planning algorithms: An extensible infrastructure for analysis and visualization," *IEEE Robotics & Automation Magazine*, vol. 22, no. 3, pp. 96–102, 2015.

[22] P. Sotiropoulos, M. A. Roa, M. F. Martins, W. Fried, H. Mnyusiwalla, P. Triantafyllou, and G. Deacon, "A benchmarking framework for systematic evaluation of compliant under-actuated soft end effectors in an industrial context," in *2018 IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pp. 280–283, IEEE, 2018.

[23] I. Garcia-Camacho, M. Lippi, M. C. Welle, H. Yin, R. Antonova, A. Varava, J. Borras, C. Torras, A. Marino, G. Alenya, *et al.*, "Benchmarking bimanual cloth manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1111–1118, 2020.

[24] A. Suarez, V. M. Vega, M. Fernandez, G. Heredia, and A. Ollero, "Benchmarks for aerial manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2650–2657, 2020.

[25] J. Leitner, A. W. Tow, N. Sünderhauf, J. E. Dean, J. W. Durham, M. Cooper, M. Eich, C. Lehnert, R. Mangels, and C. McCool, "The ACRV picking benchmark: A robotic shelf picking benchmark to foster reproducible research," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4705–4712, IEEE, 2017.

[26] R. E. Kirk, *Experimental design: Procedures for the behavioral sciences*. Sage Publications, 2012.

[27] A. Agresti, *Analysis of ordinal categorical data*, vol. 656. John Wiley & Sons, 2010.

[28] A. Makhal, F. Thomas, and A. P. Gracia, "Grasping unknown objects in clutter by superquadric representation," in *2018 Second IEEE International Conference on Robotic Computing (IRC)*, pp. 292–299, IEEE, 2018.

[29] T. Suzuki and T. Oka, "Grasping of unknown objects on a planar surface using a single depth image," in *Advanced Intelligent Mechatronics (AIM), 2016 IEEE International Conference on*, pp. 572–577, IEEE, 2016.

[30] P. Ni, W. Zhang, X. Zhu, and Q. Cao, "PointNet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3619–3625, IEEE, 2020.

[31] W. Wei, Y. Luo, F. Li, G. Xu, J. Zhong, W. Li, and P. Wang, "GPR: Grasp Pose Refinement Network for Cluttered Scenes," *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[32] 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.

[33] Q. Lei, J. Meijer, and M. Wisse, "Fast c-shape grasping for unknown objects," in *Advanced Intelligent Mechatronics (AIM), 2016 IEEE International Conference on*, pp. 509–516, IEEE, 2017.

[34] B. Denoun, B. Leon, M. Hansard, and L. Jamone, "Grasping robot integration and prototyping: The grip software framework," *IEEE Robotics & Automation Magazine*, vol. 28, no. 2, pp. 101–111, 2021.

[35] H. Dang and P. K. Allen, "Learning grasp stability," in *2012 IEEE International Conference on Robotics and Automation*, pp. 2392–2397, IEEE, 2012.

[36] Y. Bekiroglu, J. Laaksonen, J. A. Jorgensen, V. Kyrki, and D. Kragic, "Assessing grasp stability based on learning and haptic data," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 616–629, 2011.

[37] V. Ortenzi, M. Controzzi, F. Cini, J. Leitner, M. Bianchi, M. A. Roa, and P. Corke, "Robotic manipulation and the role of the task in the metric of success," *Nature Machine Intelligence*, vol. 1, no. 8, pp. 340–346, 2019.

[38] D. Shah and L. Madden, "Nonparametric analysis of ordinal data in designed factorial experiments," *Phytopathology*, vol. 94, no. 1, pp. 33–43, 2004.

[39] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[40] G. Tutz, *Regression for categorical data*, vol. 34. Cambridge University Press, 2011.

[41] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 42, no. 2, pp. 109–127, 1980.