Deep Learning Approaches to Grasp Synthesis: A Review

Rhys Newbury[1,2], Morris Gu[1], Lachlan Chumbley[1], Arsalan Mousavian[3], Clemens Eppner[3], Jürgen Leitner[4], Jeannette Bohg[5], Antonio Morales[6], Tamim Asfour[7], Danica Kragic[8], Dieter Fox[3,9], Akansel Cosgun[1]

***Abstract*—Grasping is the process of picking an object by applying forces and torques at a set of contacts. Recent advances in deep-learning methods have allowed rapid progress in robotic object grasping. We systematically surveyed the publications over the last decade, with a particular interest in grasping an object using all 6 degrees of freedom of the end-effector pose. Our review found four common methodologies for robotic grasping: sampling-based approaches, direct regression, reinforcement learning, and exemplar approaches. Furthermore, we found two 'supporting methods' around grasping that use deep-learning to support the grasping process, shape approximation, and affordances. We have distilled the publications found in this systematic review (85 papers) into ten key takeaways we consider crucial for future robotic grasping and manipulation research. An online version of the survey is available at https://rhys-newbury.github.io/projects/6dof/**

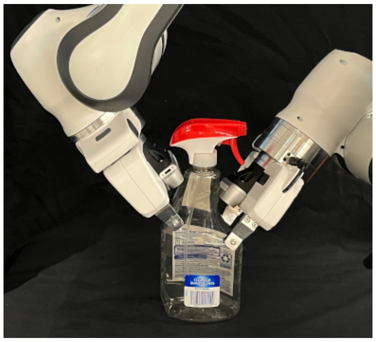


Fig. 1: Grasp Synthesis is the problem of creating a grasp pose, or a set of grasp poses. Robotic manipulators can grasp objects from multiple angles. This systematic survey reviews deep learning approaches to grasp synthesis that generate grasps utilizing all 6DoF.

I. Introduction

The manipulation of objects is a crucial skill for robotics to complete tasks in the real-world, with grasping being an integral part of such manipulation tasks. Grasping is the process of restraining an object's motion in a desired way by applying forces and torques at a set of contacts. It is an essential ability required for the majority of object manipulation tasks. Synthesizing a good grasp proposal for a specific gripper and an object or a scene made of objects is a high-dimensional search or optimization problem as there are many relative gripper-object poses, joint configurations, and contact conditions. The quality of each of these *grasp hypotheses* can be evaluated under a variety of criteria such as grasp stability, whose value depends on, for example, object or scene geometry, gripper geometry, and kinematics, as well as suitability for a specific manipulation task. Reflecting on more than four decades of research in robotic grasping, we see a change in how grasping is formulated and studied.

Early work on robotic grasping developed a theoretical framework that forms the basis of analytical approaches toward grasping [1]. At the core of this framework are contact models, which are typically based on point contacts that define what components of contact forces and torques (i.e. wrenches) can be transmitted at a specific contact and act on the object. In this framework, a *grasp* is defined as the set of wrenches that can be achieved on an object. The goal of grasp synthesis is then often framed [2] as finding a grasp that keeps the object in equilibrium in the presence of disturbances (i.e. *fixturing*) or moves it in a specific way (i.e. *dexterous manipulation*). Bicchi and Kumar [2] also mention enveloping grasps that wrap the fingers and the palm around the object, achieving more restraining grasps (i.e. *power grasps*). However, the limitation of these analytical approaches is that they assume full knowledge of object shape and geometry, material properties, and dynamics parameters. In reality, this information is rarely directly observable but can only be inferred from partial, noisy sensory data.

The increased application of data-driven approaches to computer vision has successfully transferred to robotic grasping [3], primarily addressing the complexity and uncertainty in visual perception. To better fit these computer vision techniques, the focus of robotic grasping shifted from concepts around multi-fingered, contact-based representation to pose-based ones. Commonly abstracting the robot's end-effector as an ideal two-fingered gripper approaching the object "top-down". That way, a grasp is parameterized by the position and orientation of the coordinate frame attached to the gripper or the robot wrist. Before that, the degrees of freedom of a grasp were attributed to the robot hand, its kinematic structure, and the ability to control finger movements. This simplification of grasp parameterization is further supported by the increased availability of robust and simple end-effectors - e.g. parallel jaw grippers, suction cups [4] - and under-actuated or soft hands [5, 6] that simplify the control, compared to a dexterous hand.

The increasing popularity of deep-learning has allowed the community to make significant progress over the recent years. The ability to directly synthesize grasp proposals for complex visual information allows for robots to grasp more effectively than ever before. With the speed and accuracy to find high quality grasps directly from visual information, robotic grasp-

[1] Monash University, Australia
[2] The Australian National University, Australia
[3] NVIDIA Corporation, USA
[4] LYRO Robotics Pty Ltd, Australia
[5] Stanford University, USA
[6] Jaume I University, Spain
[7] Karlsruhe Institute of Technology, Germany
[8] KTH Royal Institute of Technology, Sweden
[9] University of Washington, USA

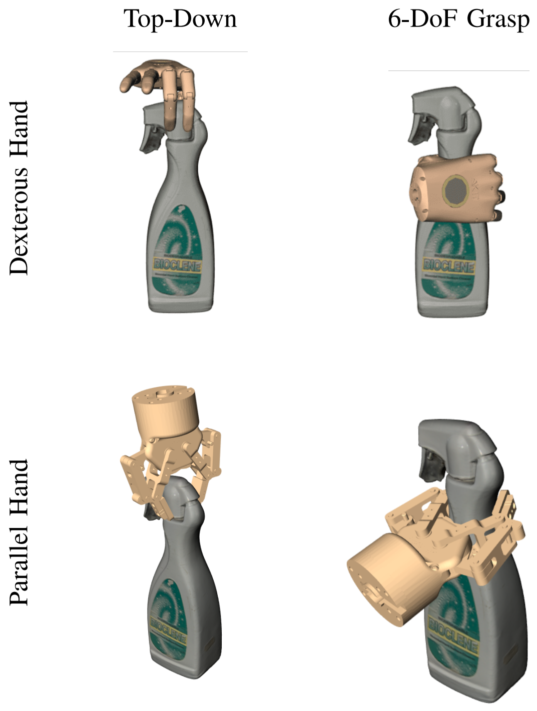


Fig. 2: This review focuses on the synthesis of 6 degrees-of-freedom (DoF) grasp pose hypotheses, where the DoF refers to the generated hand/wrist poses defined by the 3D position and orientation of a gripper specific coordinate system. This includes grasp synthesis with parallel grippers and dexterous hands alike.

ing can now be applied to highly cluttered situations [7], dynamic targets [8], complex industrial settings, (such as robotic harvesting [9]), as well as being able to grasp objects to simplify post-grasp manipulation [10].

For the purpose of this review, we reuse terms such as four and six degrees of freedom grasps, where the former encodes position in 3D and single rotation of a coordinate system attached to the gripper, relevant for defining "top-down" grasps. Six degrees of freedom grasps relate to approaches that consider the full pose of that coordinate system. Fig. 2 shows examples of both 6-DoF and 4-DoG grasps, for both parallel jaw grippers and dexterous hands. In this review, the emphasis is on deep-learning approaches applied to 6-DoF grasping, focusing on the grasp synthesis stage of the grasping process, as shown in Fig. 3. We point to other recent reviews on learning [11], point-cloud based grasping [12] and manipulation [13] for further details on the relevant areas. We believe the 6-DoF grasping is a crucial area of research for future progress in robotic manipulation. This survey aims to review the recent progress in this area systematically.

Our systematic review is based on 85 works that employ deep-learning methods for grasp synthesis, clustering the work along common methodologies, data sets, and object-sets used. From the methods' viewpoint, we devise a taxonomy along: Grasp Sampling methods, Direct Regression methods, methods incorporating Reinforcement Learning (RL), and Exemplar methods. Furthermore, we identify methods commonly employed in grasping, such as Shape Completion and Affordances. Fig. 5 depicts a visual representation of the structure of the survey paper.

The contributions of this survey are:

- A systematic review of 85 papers, focusing on deep-learning based 6-DoF grasping.
- The synthesis of the papers into 10 key takeaways (discussed in Section VII) which we consider crucial for future research in robotics and manipulation.

II. Notations and Analytical Grasping

To provide a nuanced discussion regarding the contributions of deep-learning based approaches to grasp synthesis, we briefly overview the necessary fundamentals of grasping and review notations and definitions used to define it.

- A **Grasp Pose** defines the position and orientation of a grasp. Our survey found many different formulations of grasping but most aim to learn the final pose of the robot to generate a successful grasp.
- A **4-DoF grasp** defines a grasp where hand poses are generated and defined by a 3D position and hand orientation about an approach vector that is commonly aligned with the direction of gravity and is therefore often referred to as "top-down grasping". It is often denoted by x, y, z, yaw.
- A **6-DoF grasp** defines a grasp where hand poses are generated and defined by a 3D position and orientation, thus 6-DOF in total. The major difference to 4DoF grasps is a non-fixed approach vector, providing extra flexibility but increased complexity.
- **Affordances** refers to the different tasks which can be achieved with an object [14]. This definition has been adopted in previous grasping works (e.g [15]).

An alternative way to frame a grasp, introduced by [16], is using the following three terms:

- An **approach vector** defines the line along which the gripper or robotic hand approach the target object.
- The **grasp center point** is the point in space somewhere along the approach vector where the coordinate frame fixed to the gripper must be positioned before starting to close the fingers.
- The **hand orientation** defines how the robot hand is oriented around the approach vector when placed on the grasp center point.

Some important terms used throughout the paper while describing approaches to grasping.

- Each point force applied at a contact point on the object surface also generates a torque on the object. A **wrench** summarizes this pair of force and torque applied to the object through a contact in a six-dimensional vector.
- A grasp is said to be in **Force Closure** if the forces that can be applied at the set of *frictional* contacts are sufficient to compensate for *any* external wrench applied to the object [17].
- **Hand posture** describes the configuration of the gripper or hand fingers when the grasp is started or all the contacts are made
- A **Power Grasp** is a grasp where there are multiple points of contact between the object and the fingers and palm.

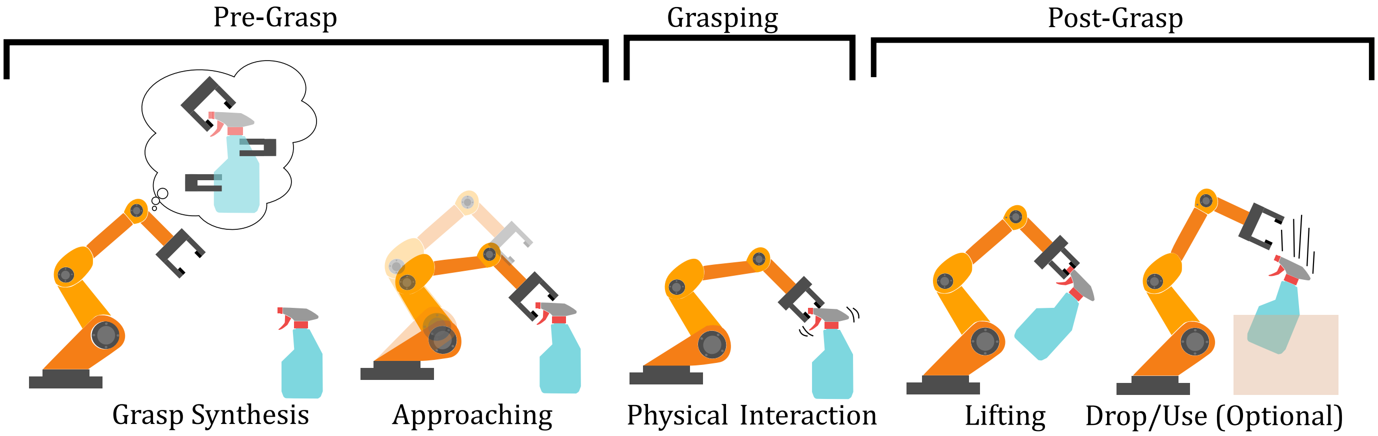


Fig. 3: Typical stages for grasping an object. Our review focuses on grasp synthesis, the first stage in the grasping process.

It maximizes the load carrying ability of a grasp and is highly stable as the enveloping nature of the grasp provides form closure [18].

- **Antipodal points** are pairs of points on the object surface whose normal vectors are collinear and pointing in the opposite direction [19]. With appropriate finger contact conditions, antipodal point grasps guarantee force closure.
- An **Antipodal grasp** is defined for two-fingered grippers that makes contact with the object at antipodal points [19].

For consistency and a thorough discussion, we provide a short insight on the relevance of analytical approaches.

The majority of methods up to the year 2000 modeled grasping analytically [2, 3]. The focus was on modeling and estimating physical conditions of grasps, such as, for example, grasp stability. A force-closure grasp was often equated with a stable grasp, although force-closure is a necessary but insufficient condition for a stable grasp [2, 20]. Physical conditions were usually simplified through approximations such as point contact models, Coulomb friction, and rigid body dynamics [3, 20, 21]. Analytic approaches have been attributed to being complex and not applicable in real-time applications. However, analytic approaches address properties of grasps, while most of the recent methods for grasp synthesis focus on positioning the hand. The advantages of analytical methods are mathematical guarantees on grasp properties, such as force-closure. This makes it easier to assess the conditions of grasps when objects are manipulated after a grasp has been applied. For example, what forces or torques can be exerted on the object before slippage occurs or how an object can be moved using in-hand manipulation. Within the first decade of the 21st century, there has been a rise of data-driven approaches to grasp synthesis [3], thanks to the development of grasping simulators such as GraspIt! [22]. Early approaches often used hand-designed features that corresponded to parts of objects that could be grasped [23–25]. Kamon et al. [26] presented one of the earliest works in 4-DoF grasping that use machine learning (ML) for grasping objects. The authors hand designed a low-dimensional feature space to estimate the quality of grasps. Since then, many works have employed traditional learning methods with a larger feature space [27–31]. One key takeaway from this early research was that while grasping can be done using RGB data only [29], depth information (RGB-D) improved grasping success [27, 30]. We refer to [3] for a more in-depth review of earlier methods for data-driven grasp synthesis. In this review, we focus on deep learning techniques in particular.

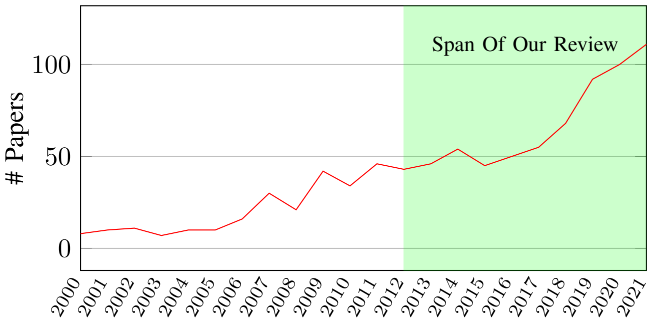


Fig. 4: The number of publications on IEEExplore that includes the keyword "Grasping" in metadata and "6DoF" in the full text is increasing year-by-year. We consider works published after Jan 1, 2012 – when AlexNet [32] was published – in our review.

III. Deep Learning Methods in 6-DoF Grasping

The number of publications investigating deep learning approaches for 6-DoF grasping have grown significantly in the last few years, as highlighted in Fig. 4. From the systematic review, we identify four main algorithmic methodologies for grasp synthesis using deep-learning based on the reviewed literature: grasp pose sampling, regressing grasp pose directly, reinforcement learning and exemplar methods. These procedures relate to how the grasps are generated at test time.

Sampling approaches consider one or many grasp samples and have learned a function to estimate the quality of a sampled grasp. An essential characteristic of sampling approaches is that each sample is evaluated *individually*. Alternatively, direct regression considers the data *globally* and learns a function to predict high-quality grasps. *RL* includes methods that involve the maximization of a cumulative reward function

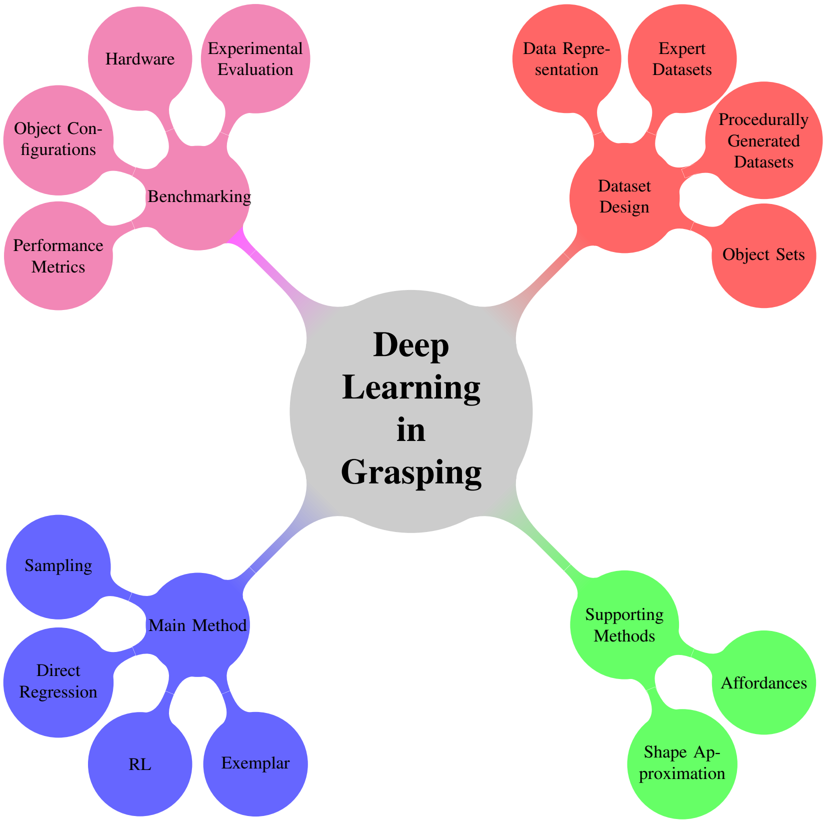


Fig. 5: A visual representation of the core topics covered in this systematic survey.

based on a robot's actions or demonstrations. Exemplar methods aim to use a similarity metric between grasps to retrieve high-quality grasps from an existing database which are most similar. These categories are often mutually exclusive, though Aktas et al. [33] employed both direct regression and sampling while Mahler et al. [34] employed both sampling and exemplar methods.

A. Sampling

We define sampling methods as any approach that considers each sample *individually* and use information encoded about the sample to make decisions about the grasp. The samples may be sourced from any discrete or continuous n-dimensional space. All reviewed sampling works are shown in Table I.

Deep-learning approaches employing sampling implement the following steps: sample information; evaluate the sample according to a quality estimation function, parameterized by a deep neural network; and optionally refine the sample using an optimization-based approach to achieve a higher quality grasp. We adopt the term 'quality', however, there is no current consensus on the definition. Throughout this paper, we use this term to refer to the confidence of grasp success. Table I presents an overview of all the *Sampling* papers. A popular deep-learning based approach for 4-DOF grasping generates a series of antipodal grasps through sampling while using a neural network to predict grasp quality [66]. This aims to predict the probability of a successful grasp to generate a series of antipodal grasps.

1) Generating Samples:

Authors can sample a subset of grasp parameters in an n-dimensional space is commonly done using one of two approaches. Random sampling occurs when samples are taken from an arbitrary random distribution. The most common distribution is a uniform distribution, however, some authors sample from other distributions such as Gaussian [48, 67]. The second method is where samples are taken at equispaced intervals within the space.

The subset of grasp parameters can be sampled from: Euclidean space, priors, configuration space, latent space, or multiple views. When sampling through Euclidean space, heuristic-based rules are often used to remove irrelevant grasp candidates.

a) Euclidean Space Sampling:

One of the most common approaches is to sample a 3D vector representing the translation part of the grasp pose. This is often achieved by sampling points from a point cloud [37, 50, 56]. The approach vector can be estimated by either using normal information estimated from the sensor data or sampling angles. ten Pas et al. [38] and Riedlinger et al. [46] use the opposite direction of the surface normal to find the grasp approach vector. Riedlinger et al. [46] uses a series of local augmentations on the initial grasp approach vector to generate a sample set of candidate approach vectors. Gualtieri et al. [37] generate candidate grasp approach vectors using the surface normal and curvature axis to generate equispaced orientations orthogonal to the curvature axis.

Year	Paper	Sample Space	Evaluation	Refinement
2015	Varley et al. [35]	Priors	Metric	
2015	Kappler et al. [36]	*Not Specified*	Binary	
2016	Gualtieri et al. [37]	Euclidian	Binary	
2017	ten Pas et al. [38]	Euclidian	Binary	
2017	Zhou and Hauser [39]	Euclidian	Binary	✓
2018	Yan et al. [40]	*Not Specified*	Binary	✓
2019	Mousavian et al. [41]	Latent	Binary	✓
2019	Liang et al. [42]	Priors	Metric	
2019	Lu et al. [43]	Euclidian	Binary	✓
2019	Ottenhaus et al. [44]	Priors	Metric	
2019	Gonçalves and Lima [45]	Euclidian	Binary	
2019	Aktas et al. [33]	Priors	Binary	✓
2020	Riedlinger et al. [46]	Euclidian	Binary	
2020	Murali et al. [47]	Latent	Metric	✓
2020	Van der Merwe et al. [48]	Hand Posture	Binary	✓
2020	Lundell et al. [49]	Multiple Views	Binary	
2020	Lou et al. [50]	Euclidian	Metric	
2020	Choi et al. [51]	Multiple Views	Binary	
2020	Lu et al. [52]	Hand Posture	Binary	✓
2020	Schaub and Schöttl [53]	Multiple Views	Binary	
2020	Murali et al. [10]	Priors	Metric	
2020	Kokic et al. [54]	Euclidian	Metric	
2021	Lundell et al. [55]	Euclidian	Metric	
2021	Lou et al. [56]	Euclidian	Metric	
2021	Lundell et al. [57]	Euclidian	Metric	
2021	Peng et al. [58]	*Not Specified*	Metric	
2021	Jiang et al. [59]	Euclidian	Binary	
2021	Kasaei and Kasaei [60]	Multiple Views	Binary	
2021	Wang et al. [61]	Priors	Binary	
2021	Munoz [62]	Multiple Views	Binary	
2021	Corsaro et al. [63]	Priors	Binary	
2021	Ren et al. [64]	*Not Specified*	Metric	
2021	Wen et al. [65]	*Not Specified*	Metric	

TABLE I: Publications with deep-learning focused sampling methods. We cluster the papers based on the space the sample through and how the samples are evaluated. Some approaches further consider an optional refinement stage.

Some approaches sample angles independently of the surface normals. Lou et al. [50, 56] chose N points from a point cloud and sample the wrist angles randomly for the grasp. However, they restrict the approaching vector of the grasp to be above the table. Kokic et al. [54] randomly sample grasp and roll angles, and offset distances for each point in the point cloud. Both Lu et al. [43], Lundell et al. [55, 57] sample grasp candidates around the center of an object with a random orientation.

Sampling can also be performed with regularly spaced points through euclidean space. Jiang et al. [59] sample regularly spaced points for the position of the grasp pose. Similarly, Gonçalves and Lima [45] sample equispaced points throughout the region of interest.

Alternatively, instead of sampling a xyz vector, angles can be used. Zhou and Hauser [39] chose a random hand orientation and approach vector after which they translate the hand along the approach vector of each sample until a grasp is found that does not collide with the gripper when it is open but collides when closed.

In Euclidian space, sampled grasp are commonly pruned to remove infeasible grasps based on rules such as:

- The robot hand is in collision with the point cloud when fingers are open [38, 39].
- The closing region of the fingers does not contain at least one point from the point cloud [37–39].

b) Sampling Priors: More complex algorithms can also be used to find a set of feasible grasp candidates. These are often used when creating grasp samples which will then be evaluated for affordances [10] or different types of grasps (e.g power, pinch grasps) [63]. The most commonly [10, 42, 61, 63] used grasping algorithm is an SVM-based approach proposed by Pas and Platt [68]. Mahler et al. [34] use a modification of the grasping algorithm proposed by Smith et al. [69] to generate a series of antipodal grasps. They frame the problem as a multi-armed bandit, to be solved using Thompson sampling [70].

While grasping simulators are often used for generating training data, they have also been used at test time to sample grasps. A common prerequisite for grasping simulators is a full approximation of the shape model. The common simulators for this purpose were GraspIt! [22] or Simox [71]. Ottenhaus et al. [44] generated grasp samples on a reconstructed object using the grasp planner by Simox [71]. Alternatively, Varley et al. [35] sampled grasps using the Simulated Annealing planner [72] for partially visible objects.

A deep-learning approach can also be used to synthesize grasp samples. Aktas et al. [33] used the direct regression based approach from Kopicki et al. [73] to generate multiple grasps, to be used as grasp samples.

c) Latent Space Sampling: Mousavian et al. [41] train a Variational Auto Encoder [74] and uniformly sample through latent space to generate grasp poses. An example of the grasps generated by this approach is shown in Fig. 6. This sampling approach is also adopted by Murali et al. [47].

d) Hand Posture Space Sampling: Some authors sample grasp configurations from a prior distribution fit to the training set to create an initial hand configuration for their approach [48, 52, 67]. Van der Merwe et al. [48] and [67] fitted a Gaussian Mixture Model to represent a grasp prior function trained on grasp configurations seen during training. Similarly, Lu et al. [52] trained a Mixture Density Network [75] over all grasp attempts from the training set.

e) Multiple Views: A number of papers employ multiple viewpoints of the scene, where these can either be from virtual or real cameras. A camera viewpoint is then sampled, and a grasping approach is employed on the sampled viewpoint. Schaub and Schöttl [53] combine multiple viewpoints around the scene to create a 3D representation of an object. Using the 3D representation, they generate depth images for a series of virtual cameras. They then use a 4-DoF grasping algorithm [8] for each real and virtual camera. This is extended by Schaub et al. [76] who fused depth images around the scene to provide more detailed depth maps. Choi et al. [51] extended a previous 4-DoF approach [77] to 6-DoF by using an iterative improvement algorithm approach to choose an approach direction. Munoz [62] and Kasaei and Kasaei [60] both generate multiple views of the object from virtual cameras using a captured point cloud from a single viewpoint. They proposed a method to select a view according to an entropy-based measure. Lundell et al. [49] use the algorithm proposed by Satish et al. [77] on the depth map from multiple viewpoints. The robot then executes the best grasp from all viewpoints.

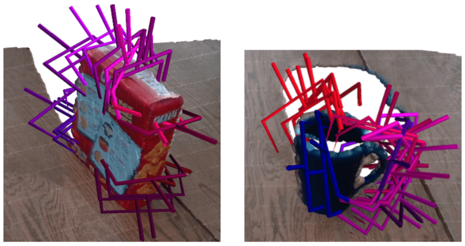


Fig. 6: Mousavian et al. [41] sample through the latent space of a trained model to generate a series of grasp candidates (©2019 IEEE)

2) Sample Evaluation: Once samples are generated, these approaches use a function, commonly parameterized by a neural network, to estimate a numerical value representing a grasp metric. This is commonly used an estimation of the 'quality' of the grasp. The grasp quality can be approximated through an analytical metric or the likelihood of a grasp being successful. While some approaches execute the highest quality sample directly, others refine the samples using optimization-based techniques based on the learned quality function.

a) Binary classification: Predicting grasp success can be treated as a binary classification problem, where the output represents the confidence of a grasp being successful or not. A CNN can be trained for the binary classification of grasp success, with the network either outputting 0 or 1 for a successful grasp [33, 36–41, 43, 45, 46, 48, 52, 61, 63]. During the collection of ground-truth training data to determine if a grasp is successful, the grasp is most commonly executed using a robotics simulation environment [33, 36, 39–41, 43, 46, 48, 52, 59, 63], where a grasp will be labeled successful if the object remains within the gripper after being lifted. Grasp success can also be based on an analytical metric, such as whether the grasp would form a force-closure grasp [38, 45] or if the grasp is antipodal [37].

The binary classification of a sampled grasps can also be consider as one of the outputs of a sampling approach. For example. Jiang et al. [59] designed an approach which learns implicit functions that predicts grasp parameters (quality, rotation and width) from the feature space representation of a randomly sampled query point. Van der Merwe et al. [48] trained a network to predict the success of a grasp, as well as a signed distance function that represents a distance between a query point and the surface of the object.

b) Learning a metric: Learning metrics associated with grasps, rather than a pure binary label, have been proposed to better represent the quality of a grasp. These metrics are often continuous numbers (rather then the previously described binary label) that provide additional information to rank grasp quality or help "guide" deep learning algorithms using them as a fitness score.

Varley et al. [35] learn a series of heatmaps, where each pixel represents the location's efficacy as a fingertip or palm location for common grasp types found in their training set. Liang et al. [42] designed an approach which learns a grasp quality metric based on the force-closure metric and wrench space analysis [78]. Similarly, Ottenhaus et al. [44] train a CNN to estimate the force-closure probability of a grasp under a small random perturbation. Lundell et al. [55] and Lundell et al. [57] trained a grasp classifier using a *Generative Adversarial Networks* (GANs), and use the discriminator loss to help produce realistic-looking grasps. Wen et al. [65] compute a continuous score for each grasp by looking at the stability of randomly sampled grasps in the proximity of the selected grasp. They argue that "grasp stability should be continuous over its 6D neighborhood", therefore, this should allow for more stable grasps. Furthermore, this can allow for imperfections in the robotic grasp position whilst still being likely to execute a stable grasp. Peng et al. [58] and Ren et al. [64] designed an approach to learn the smallest co-efficient of friction which will satisfy a force-closure grasp. Learned metrics can also represent the quality of the grasp with respect to other aspects of the task, for example, metrics based on the relevance of the grasp to a given affordance [10, 54].

Other approaches consider workspace constraints. Lou et al. [50] designed an approach to learn the probability of grasp success and use an additional network to learn the probability of a grasp being reachable. These probabilities can be multiplied to find the likelihood of success for the entire grasping action. This work is extended by Lou et al. [56] to allow the robot to grasp in constrained environments, such as in boxes, where walls may limit the success of grasping an object.

3) Optimization-based Grasp Refinement: Gradient-based optimization through the trained network can be used to find high-quality grasps candidates. A sampled grasp is taken as the initial seed, which is then refined based on the derivatives of the quality estimation network. This attempts to maximize estimated grasp quality. Zhou and Hauser [39] train a CNN to predict grasp quality given a depth image and a sampled end-effector pose. The end-effector pose is locally optimized using the quasi-newton method on the gradient of the learned quality function. Similarly, some authors take the derivative of the grasp quality with respect to the grasp pose and then use gradient ascent to refine the grasp candidates [40, 41, 43, 47]. An example of the refinement process is shown in Fig. 7, where the initial grasp is a potentially bad sample (blue) and is refined to a higher quality grasp (yellow). Both Van der Merwe et al. [48] and Lu et al. [52] treat finding a grasp configuration as an optimization problem aiming to maximize the probability of grasp success, seeding the process with samples from a prior distribution. Lu et al. [67] extended this using active learning during training, improving their results. In contrast to other works, Aktas et al. [33] found that optimization of the sample did not improve the grasp success rate in simulation. The authors explore the use of gradient-based optimization and simulated annealing for their optimization.

B. Direct Regression

Direct regression approaches process the entire sample space *simultaneously*. This could be used to predict properties of the grasp, such as grasp parameters and grasp quality from

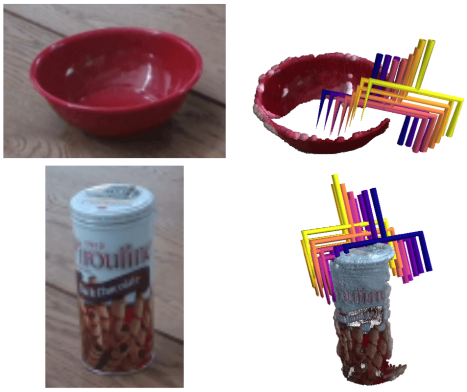


Fig. 7: An example of a grasp refinement process from [41], where the initial grasp is shown in dark blue and the final grasp pose is shown in yellow. (©2019 IEEE)

Year	Paper	Grasps	Direct Pose	Multi-stage	DimRed
2017	Veres et al. [79]	Single	✓		
2018	Schmidt et al. [80]	Single	✓		
2018	Choi et al. [81]	Single			✓
2019	Liu et al. [82]	Single	✓		
2020	Qin et al. [83]	Multiple			✓
2020	Yang et al. [84]	Single	✓		
2020	Jeng et al. [85]	Multiple			✓
2020	Wang and Lin [86]	Single			✓
2020	Fang et al. [87]	Multiple		✓	
2020	Wu et al. [88]	Multiple			✓
2020	Ni et al. [89]	Multiple			✓
2020	Breyer et al. [7]	Multiple			✓
2020	Liu et al. [90]	Single	✓		
2020	Shao et al. [91]	Single		✓	
2020	Ni et al. [92]	Multiple		✓	
2021	Wang et al. [93]	Multiple		✓	
2021	Sundermeyer et al. [94]	Multiple			✓
2021	Zhao et al. [95]	Multiple		✓	
2021	Gou et al. [96]	Multiple			✓
2021	Zhu et al. [97]	Multiple		✓	
2021	Wei et al. [98]	Multiple		✓	
2021	Li et al. [99]	Multiple			✓
2021	Li et al. [100]	Single		✓	

TABLE II: Approaches to direct regression of 6-DoF grasping using Deep Learning, either generating a single or multiple grasp poses as an output. We found three main approaches: regress to a pose directly (Pose), employing a multi-stage approach, or perform dimensionality reduction (DimRed).

visual information. Direct regression approaches are often framed as an end-to-end solution for finding grasps. Direct regression based methods attempt to reduce computational cost compared to sampling methods by processing data globally through the network once. Early works in deep-learning based direct regression are inspired by object detection work from the Computer Vision community [101, 102]. Authors would treat finding top-down grasps similar to detecting objects in an image and would use the depth image to recover the 3D position of the grasp [103–106]. Morrison et al. [8] designed a heat-map based approach for 4-DoF grasping that was designed for fast inference, enabling closed-loop grasping and grasping of moving objects. However, as the dimensionality of the output increased, the problem difficulty also increased [41, 107]. This presents scaling issues when directly regressing to high-dimensional outputs such as a 6-DoF grasp pose or a high-DoF dexterous robotic hand configuration. To overcome this, researchers often reduce the DoF of the output from a single network. This can be accomplished through multi-stage approaches, where each stage has a specific task, or by reducing the DoF by conditioning the output on an input. The direct regression approaches found in our systematic review are shown in Table II.

a) Directly Regressing a Pose: Schmidt et al. [80] presented one of the earliest works to train a CNN to directly estimate a single 6-DoF grasp pose of an end-effector from input depth images. Similarly, Yang et al. [84] trained a network to estimate the transformation matrix needed to be applied to the end-effector to produce a successful grasp. However, both approaches assume that there is only a single, most optimal ground truth grasp for each input. This would introduce ambiguities as there may potentially be numerous successful grasps that can be executed for any input. To account for this, Liu et al. [82] designed a loss function that accounts for multiple ground truth grasps in the training data by calculating the loss between the current output and the closest ground truth grasp. This is extended in another work [90] that includes differentiable terms for a grasp metric and self-collisions in the loss function. This allows their approach to work both in an unsupervised manner or using a smaller supervised dataset. Veres et al. [79] create a generative model based on a *Conditional Variational Autoencoder* (CVAE) [108]. They use the CVAE to generate contact positions and contact normals for a multi-fingered robotic hand.

b) Reducing DoF: Due to the difficulty of regressing all 6 DoF of a grasp, some DoF can be reduced by analytically determining some DoF conditioned on the regressed DoF. This is seen in some 4-DoF approaches, for example,, the depth of the grasp is not directly regressed but instead uses the depth image to recover the grasp depth conditioned on the grasp position and 2-finger gripper wrist rotation [8].

Sundermeyer et al. [94] reduced the 6-DoF grasping to a 4-DoF representation by ensuring one of the contact points for a two-finger parallel gripper was taken from the point cloud. The 3-DoF rotation and gripper width was then estimated. Breyer et al. [7] directly output the predicted grasp quality, orientation, and opening width for each voxel in a queried 3D volume. The 3D position is recovered from the center of a voxel. Jeng et al. [85] propose a coarse-to-fine representation, where the orientation is initially coarsely discretized as a grid with a given confidence value. A refinement step is also used for each grasp pose to allow further flexibility from the discretized coarse representation. Gou et al. [96] estimate an $SO(3)$ orientation and confidence of every pixel directly from an RGB image. An analytical method based on the depth image is then used to find the gripper width and position for each pixel. Ni et al. [89] directly regressed grasps from sparse point clouds. They predict the grasp quality, approaching direction, and opening direction of the gripper for every point in the original point cloud. Similarly, the approach by Li et al. [99] estimates the rotation, width, depth, and quality of a grasp for each point in the point cloud. They combine this with another branch which predicts whether the grasp would be in collision. Alternatively, Choi et al. [81] predicts the most likely grasping direction and wrist orientation from a set of discretized directions and orientations. The translational part of

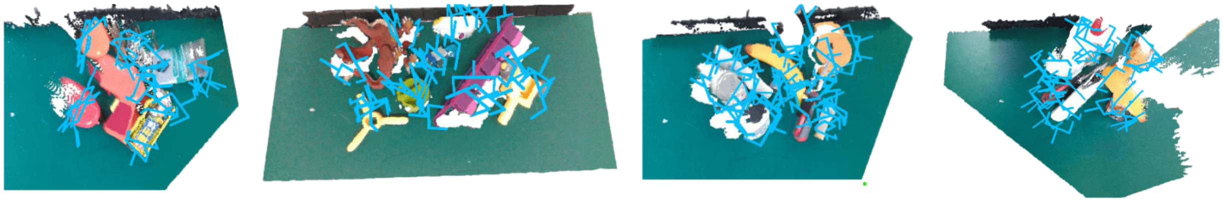


Fig. 8: Fang et al. [87] uses a direct regression approach to generate a series of grasps from real-world data. (©2020 IEEE)

the pose is found by determining the centered contacting voxel within the grasping direction. Wang and Lin [86] assume they are grasping the centroid of the object and attempt to directly regress the quaternion for the grasp. Wu et al. [88] create a network with three branches, trained to predict the likelihood of the grasp consisting of antipodal points, grasp offset for each of the sampled SE(3) points, and grasp confidence (trained on whether grasp was executed in simulation). Qin et al. [83] trains a neural network based on Qi et al. [109] to predict both grasp point and quality for every point in a point cloud.

c) Multi-Stage Approach: These approaches use multiple stages, where a stage refers to a component of the end-to-end model with a loss function and at least one specific task. This aims to simplify learning by breaking the problem down into smaller parts. Most multi-stage approaches consider three-stage approaches [87, 93, 95, 97, 100], while, Wei et al. [98] propose a two-stage approach. Stages are commonly done in series (one after another), however, the last two stages in the work by Fang et al. [87] are in parallel (simultaneously).

The first stages of the network are commonly used for at least one of the following: predicting grasp quality for subsampled points [87, 93, 95], estimating a grasp for each point [92, 98], estimating a subset of the DoF [87] or employing contrastive loss functions [97]. The middle stages can be used to create grasp proposals (if they were not generated in a previous stage) [93, 95] or further estimate a subset of the DoF [97].

The last stage can act as a refinement stage to improve the regressed grasps [92, 95, 98]. Fang et al. [87] used two final parallel stages to generate grasp proposals and predict the ability of the grasp pose to tolerate larger errors, aiming to improve the robustness to imperfect sensing or control. The final stage can also be used to predict the remaining DoF [97]. Example grasps regressed in the work by [87] from real-world data are shown in Fig. 8.

An alternative multi-stage approach is demonstrated by Shao et al. [91] and Li et al. [100] who propose an approach that uses each stage of the network to predict a single contact point. The subsequently regressed contact point will be conditioned on the previous point. The final grasp is then recovered from the regressed contact points. Furthermore, Shao et al. [91] shows that this approach is generalizable between different robotic hands.

C. Reinforcement Learning

Reinforcement Learning (RL) approaches aim to learn a policy to maximize the cumulative reward commonly over a multi-step task. We only review deep RL, where the policy is parameterized by a deep neural network. We sub-divided the reviewed works into two main approaches: On- and Off-Policy Learning. For a more comprehensive review on deep RL, see [110] and for a deeper exploration of RL in grasping and its open challenges, see [111].

Year	Paper	Learning	Algorithm	LfD
2018	Gualtieri and Platt [115]	On Policy	DQN[116]	
2019	Wu et al. [117]	On Policy	PPO[118]	
2019	Merzic et al. [119]	On Policy	TRPO [120]	
2020	Mandikal and Grauman [121]	On Policy	PPO[118]	
2020	Song et al. [122]	Off Policy	Q Learning[123]	✓
2020	Wu et al. [107]	On Policy	PPO[118]	
2021	Berscheid et al. [124]	On Policy	Single Step MDP	
2021	Kawakami et al. [125]	On Policy	PPO[118]	✓
2021	Tang et al. [126]	Off Policy	Q Learning[123]	
2021	Wang et al. [127]	Off-Policy	DDPG[128]	✓

TABLE III: Our systematic survey found 10 publications employing *Off Policy* or *On Policy* reinforcement learning (RL).

Some seminal works in reinforcement learning for grasping include the work by [112–114]. Levine et al. [112] collected 800k grasp attempts over multiple months. They present a self-supervised approach, and the authors claimed their approach is analogous to an RL formulation. Kalashnikov et al. [113] and Quillen et al. [114] presented works that formulate grasping as a reinforcement learning problem. Kalashnikov et al. [113] focus on real-world data collection, collecting 580k real-world grasping attempt, while Quillen et al. [114] train on purely simulated data. Table III shows and summarizes all reinforcement learning based works found during our review.

1) On-Policy: We found that out of the reviewed work employing RL, On-Policy methods were more common [107, 115, 117, 121, 124, 125, 127, 129]. In On-Policy RL, training a policy is done using experiences that are collected from the most recent policy. In the work from Kawakami et al. [125], the grasping task is divided into consecutive stages: orienting the end-effector, approaching the target, and closing the gripper. A different RL model is trained for each stage, and curriculum learning is employed that adjusts the reward function based on the success rate of each task. Gualtieri and Platt [115] used a Deep Q-Network [116] with Monte Carlo updates to learn how to grasp an object and place it into a desired configuration. Working with a 24-DoF hand, Mandikal and Grauman [121] trained an actor-critic model. They proposed a two-step architecture: Initially, a CNN, which is trained on ContactDB [130], estimates the pixel regions that belong to a "use" affordance. An RL policy then takes this affordance mask, RGB-D image, and gripper configuration as input and outputs the wrist pose and the 24-DoF robot hand configuration. A sparse reward is awarded when the object is lifted from the ground, and a dense reward is awarded according to a distance metric from the affordance region. Chen et al. [129] used an advantage actor-critic policy gradient to train a policy that will optimize the viewpoint for grasping. They then apply the grasping algorithm developed by Pas and Platt [68] to calculate grasps for the optimized viewpoint. Wu et al. [117] used a single depth image and introduces a novel attention mechanism that learns to focus on sub-regions of the depth image in order to grasp better in cluttered environments. They formulate the problem using a policy gradient method based on PPO [118]. Wu et al. [107] extended this framework to robotic hands with arbitrary degrees of freedom. Merzic et al. [119] propose the use of TRPO to learn control policies that take contact feedback as input. They show that this policy significant improved the robustness of the grasp under both object pose uncertainty and shape complexity.

2) Off-Policy: Off-Policy RL methods use the data collected throughout training to train a new policy. Employing human demonstrations, Song et al. [122] used Q-learning to estimate the optimal Q-function. They simulate future states by giving an action for the robot to complete, allowing the algorithm to forward simulate possible future states conditioned on the current state-action pair. Tang et al. [126] demonstrated collaborative pushing actions to facilitate grasping. Their approach uses Q-learning to learn a deterministic policy for pushing and grasping. No reward is assigned to pushing actions - the agent is only rewarded when the robot successfully grasps the object. Berscheid et al. [124] formulated the problem as a Markov Decision Process with a single action step. They train a fully convolutional neural network to learn a 4-DoF planar grasping system. A model-based controller decides the other two degrees of freedom by avoiding collisions and maximizing grasp quality. Wang et al. [127] trained a grasping policy from demonstrations based on the Deep Deterministic Policy Gradient algorithm [128]. The demonstrations are obtained using a motion and grasp planner, which is assumed to be an ‘expert’ in their formulation.

D. *Exemplar Methods*

Exemplar methods attempt to transfer grasps from previous examples. Patten et al. [131] designed an approach to grasp novel objects by learning from experience. This is achieved using metric learning to encode objects with similar geometries nearby in feature space. Finding a successful grasp is framed as a nearest neighbor search through feature space, searching for a previously successful grasp. The approach by Mahler et al. [34] compared a given grasp candidate against a database of successful grasps. They compare grasp candidates using three feature maps: grasp parameters, depth match gradient at local patches around the object, and similarity of the object model assessed by a deep-learning based network. The feature maps form a prior belief distribution on the similarity to all grasps in the database. The grasp with the maximum lower confidence bound of the belief distribution is executed.

IV. Supporting Methods Based on Deep Learning

Deep-learning can be applied to methods throughout the grasping pipeline that aim to improve the success of a grasping task. This task does not have to focus solely on picking up an object. More complex manipulation tasks may require grasping that affords a particular subsequent action. For example, when handing over a full mug, the robot may need to grasp the mug handle. We found two clusters of supporting methods in the reviewed literature: shape approximation techniques and affordances based methodologies.

A. *Shape Approximation*

The most common form to approximate the shape of an object from partial information found in the literature is *Shape Completion*, which aims to estimate the full object model from a partial input shape (e.g. point cloud of an object from one camera view). We define *Shape approximation* more generally to include any method which approximates shape from an input. This includes the approximation of the actual shape of an object by simple(r) shapes and the fusion of multimodal data to approximate a shape.

1) Shape Completion: Varley et al. [132] trained a 3D CNN to employ shape completion on a single view voxel grid, outputting a voxel grid with shape completed object. Gao and Tedrake [133] instead use the method from Zhang et al. [134] for shape completion, predicting a 3D voxel grid directly from RGB-D images. Kiatos et al. [135] use a variational autoencoder [136] to predict the occluded surface points and associated normals of a partial 3D point cloud. Chavan-Dafle et al. [137] predicted the depth image that estimates the ‘back’ side of an object from a masked depth image. The front and back sides can then be stitched together quickly to form an object mesh.

The uncertainty around the output of the shape completion can be useful. Gualtieri and Platt [138] incorporate uncertainty in their shape completion network, where it represents the estimated probability that each predicted point is accurate. Another approach including uncertainty is Lundell et al. [139] which incorporates a Monte-Carlo drop-out procedure [140] to generate a series of shape completed objects. GraspIt! [22] is then used to plan grasps over the mean object shape. The most suitable grasp over the series of shapes is chosen as the grasp point. An example of this procedure is shown in Fig. 9. Interestingly, uncertainty is ignored in follow-on work from the same authors [49, 55, 57].

2) Auxiliary Tasks: In deep learning, auxiliary tasks can be added into a deep-learning model, which has been shown to boost the performance of the model in some domains [141]. Jiang et al. [59] asserted that 3D reconstruction and grasping are closely related, where both relies on knowledge of an object's local geometries. Authors have proposed learn object reconstruction as an auxiliary task to grasping [40, 59, 84]. Jiang et al. [59] used a self-supervised approach to reconstruct an object and calculate a grasp. Yang et al. [84] simultaneously regressed a grasp pose and reconstructed the point cloud of an object. The grasp pose is then refined by projecting it onto the surface of the point cloud reconstruction. Yan et al. [40] employed two networks, one for shape completion and another for grasp outcome prediction. They demonstrated a performance improvement when the grasping network uses the feature space representation produced by the shape generation network.

3) Other: Avigal et al. [142] do not complete explicit shape prediction, however, they used a network that takes RGB images from multiple viewpoints and generates the corresponding depth maps for the shape. The depth maps are then fed into a 4-DoF grasping algorithm [143]. Ottenhaus et al. [44] used Gaussian Process Implicit Surfaces [144] to fuse visual and tactile sensor inputs. After capturing a point cloud of the object, the robot gathers information about unseen sides of objects using tactile information. Researchers have shown that many everyday objects can be modeled as simple shapes [145, 146]. Using this observation, Torii and Hashimoto [147] approximated objects as a series of 3D primitive shapes (hexahedron, cylinder, sphere). They use a neural network to predict the likelihood of each primitive shape before using a pre-computed database of rules to perform the grasping.

B. Affordances

This section reviews how affordances have been used in the domain of 6-DoF grasping. See [148] for a review of affordances in the more general robotics domain. In addition to considering the success of grasping, these approaches have additional considerations for what kind of task it is used for. The robot's understanding of how an object is used in a particular task or how humans use those objects can lead to higher-level reasoning about the grasping task. This lends itself towards 6-DoF grasping, as different parts of the object need to be grasped in different ways, depending on location of the affordance.

Some researchers use deep neural networks to segment objects for different affordances [121, 149–152]. They then use analytical methods to find grasps within the segmented object portion with an appropriate affordance. Alternatively, the approach by Murali et al. [10] estimated the quality of a sampled grasp given an affordance label. This is shown in Fig. 10, where the grasps shown in green are relevant to the given affordance. Ardón et al. [15] created a knowledge base graph representation using Markov Logic Networks to obtain a probability distrib ution of grasp affordances. Additionally, both Manuelli et al. [153] and Gao and Tedrake [133] presented an approach based on detecting a fixed number of keypoints for a category of objects.

Object Set/Database	# of times used	Sim//Real
YCB [154]	29	Real
3DNET [158]	15	Sim
BigBIRD [155]	12	Real
KIT [156]	12	Real
ShapeNet [157]	11	Sim
Grasp [36]	10	Sim
EGAD! [162]	2	Sim/Real
Cornell [104]	2	Real
Dex-Net [143]	2	Real
PSB [159]	1	Sim
ModelNet [160]	1	Sim
ObjectNet3D [161]	1	Sim
ContactDB [130]	1	Sim
Procedural [163]	1	Sim
Custom	11	Real/Sim

TABLE IV: Number of times objects sets are used in reviewed papers.

V. Dataset Design

A. Objects Sets

The objects used for training and testing grasping algorithms are crucial for the reported grasp success and allowing the community to reproduce the results. Researchers commonly use subsets of existing object sets when investigating grasping. However, there is no standard procedure for selecting this subset. This can lead to inconsistencies in objects between different works which are using the same object set. This increases the difficulty of comparing grasping performance between works. The most commonly used object set in the reviewed works is Yale-CMU-Berkeley(YCB) [154], being used almost twice as often than the next most adopted object set (Table IV).

YCB [154] (shown in Fig. 11), BigBIRD [155], KIT [156] and Cornell [104] consist of mostly household items such as food, toys and tools. These object sets are appropriate for service robotics, but may not test the robustness of grasping algorithms on complex objects. ShapeNet [157], 3DNet [158], Grasp [36], PSB [159], ModelNet [160], ObjectNet3D [161] and ContactDB [130] include object model repositories, containing a large number of virtual object models, mainly used for training and testing in simulation.

EGAD! [162] and Procedural [163] consist of procedurally generated object models. EGAD! [162] proposes a set of 3D printable objects that vary in terms of grasping difficulty [143] and object complexity [164]. Procedural [163] generate a simulated object set by attaching rectangular prisms in random orientations and locations.

Most real-world object sets do not provide a standardized method to acquire the physical objects consistently (an exception to this is YCB[1]). One solution to this is 3D printed datasets such as EGAD![2], however, these objects lack semantic meaning.

B. Procedurally Generated Datasets

Even though a majority of the reviewed works use benchmark object sets, they commonly opt to create their own

[1] https://www.ycbbenchmarks.com/
[2] https://dougsm.github.io/egad/

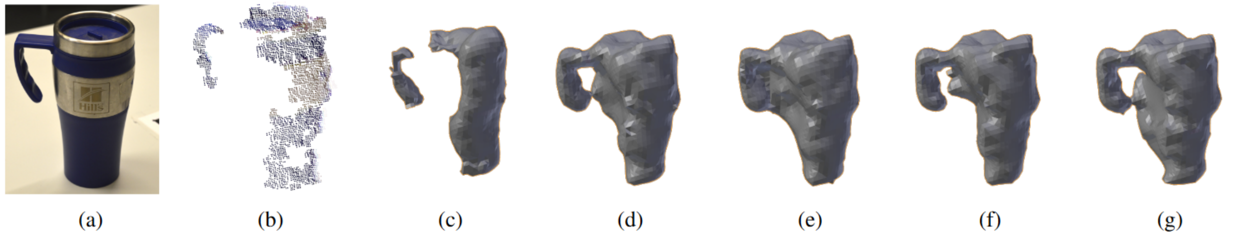


Fig. 9: Lundell et al. [139] uses drop out layers to generate 20 shape completed samples, shown in (e) - (g). The average of these samples is shown in (d) compared to [132](c). (©2019 IEEE)

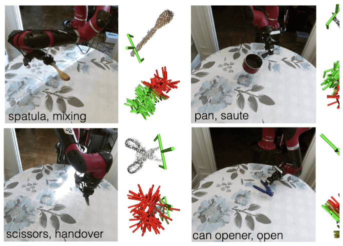


Fig. 10: Murali et al. [10] studies task-oriented grasps on unknown objects. For each sample, the top visualization shows the grasp with the highest quality considering a given affordance. The bottom shows all the stable grasp candidates, colored by the relevance to the affordance (green is high)

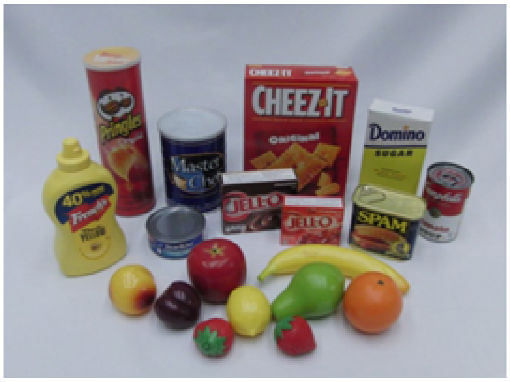


Fig. 11: A commonly used object sets in robotic grasping is the YCB object set. It consists of a set of daily household objects, with a subset of the objects from the 'Food' category shown here. (©2015 IEEE)

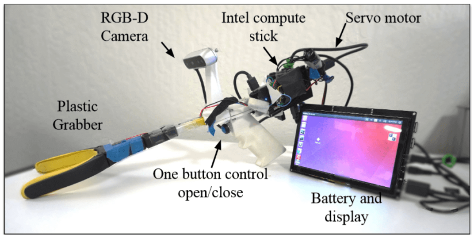


Fig. 12: Song et al. [122] design a low-cost handheld gripper to generate human annotations for grasping. (©2020 IEEE)

custom datasets with those objects. A number of works used datasets collected using real robots [60, 61, 81, 96, 124] and some combined simulation and real-world data [61, 87, 96]. However, the large majority opted to use purely simulated datasets when training their networks.

Some authors have released their datasets, with reviewed works employing public datasets such as GraspNet-1Billion [87], a hybrid 6-DoF grasping dataset that captures real RGB-D camera data and combines this with simulated grasp poses, Shape Completion Grasping [132], a database of voxel grid pairs for shape completion and ACRONYM [165], a simulation-based dataset for 6-DoF grasping.

C. Expert Datasets

Researches have proposed various datasets which consist of expert demonstrations, either from a human or an algorithm. Yan et al. [40] generates a dataset of around 1.6k human grasping demonstrations within Virtual Reality (VR) from 5 people. Also in VR, Kawakami et al. [125] creates a system to collect grasping demonstrations. The operator demonstrates the position of the arm using a controller with tracked position and pose. Using a handheld gripper (see Fig. 12), Song et al. [122] generated a dataset from human demonstration. The dataset contains 12 hours of gripper-centric RGB-D videos, with each picking attempt separated into short clips to correspond to grasps. Similarly, Diaz Cortes et al. [166] generated 300 demonstrations by kinesthetically teaching the robot trajectories. Osorio et al. [167] created a dataset of human grasps where the human is controlling a robotic gripper with a joystick. Alternatively, the approach by Wang et al. [127] used demonstration from Optimization-based Motion and Grasp Planner (OMG planner) [168]. Taheri et al. [169] presented a dataset of whole-body grasps generated by ten subjects interacting with 51 everyday objects of varying shape and size. The object meshes are annotated with the contacts created by the human hand.

D. Data Representation

There are four major sensor inputs used for the deep learning methods: Point Cloud; Voxel Grid; RGB-D Image; and Depth Image. These four representations are all interchangeable for spatial data, assuming that intrinsic parameters for the camera are known. Table V shows the popularity of each of the input formats and Table VI shows the popularity of different network architectures.

Point Cloud

Point clouds are the most popular data format, especially with the advancements in networks that can learn directly from point clouds, such as PointNet [170]. In addition, the point cloud representation allows researchers to easily fuse data from multiple viewpoints if the relative camera poses are known.

Input Format	Number of times used
Point Cloud	22
Depth Image	15
RGB-D Image	12
Voxel Grid	10
Segmentation Mask	9
Other	10

TABLE V: Number of times each input format has been used

For point cloud based data, researchers commonly used PointNet [170] and PointNet++ [109]. This network backbone type has found uses in different methods including direct regression [83, 85, 87–89, 94, 95], RL [127], sampling [10, 41, 42, 47, 58] and shape completion [138].

Point cloud for direct regression tends to subsample points and learn grasps for them [54, 61, 83, 85–87, 89, 92, 95, 98–100]. When a point cloud is used for sampling, one of the following procedures is usually followed. The points inside the gripper are fed to the network [42, 63, 64], the points will be transformed such that the origin aligns with the grasp frame [63, 65], or a representation of the gripper will be rendered, and all of the points will be fed into the network [41]. Minimal surveyed work has investigated the use of point cloud in RL apart from [127].

Images

The next most popular formats are depth and RGB-D images. Learning from images has been highly studied in both computer vision and robotics, allowing researchers to train pre-existing architectures for robotic grasp synthesis.

For images, the most common approach is to employ ResNet [171] as the backbone [55, 57, 84, 90, 96, 97, 122, 131, 172]. In addition to ResNet, various other architectures have been used, including VGG [173] LeNet [174], DenseNet [175] and U-Net [176].

Direct regression approaches commonly use the whole image as input [80, 96, 97]. For sampling, some approaches generated a depth image slice representing points within the gripper [37, 38] while others concatenated grasp features after processing images through a CNN [33, 43, 55, 57]. RL approaches input the current camera view to generate an action [107, 115, 117, 122, 124, 126, 166].

Voxel Grids

A voxel represents a value in a regular 3D grid. Voxel Grids are analogous to images in 3D space, where a voxel is similar to a pixel over a 3D grid instead of a 2D image.

For voxel based inputs, VoxNet [177] is most commonly used. VoxNet integrates a volumetric occupancy grid representation of the data with a 3D-CNN. Voxel grids have been used for shape completion [132, 135], sampling [50, 52, 67] and direct regression [7, 81, 82]. For sampling, Voxel grids have been used similarly to point clouds, transforming the origin to align with grasp frame coordinates [50] or adding grasp configuration features after the convolutional layers [44, 52, 67]. Voxel grids have also been used to directly regress a single pose using a subset of the voxel grid corresponding to the object [81, 82] as well as directly regressing a grasp for each voxel [7].

Network Backbone	Number of times used
PointNet/PointNet++ [109]	14
ResNet [171]	6
LeNet [174]	3
Other	14
Custom	26

TABLE VI: Number of times each network backbone has been used

Category	Popularity
Robot Arm	66
Humanoid	6
Mobile Arm	3
Not Used	7

TABLE VII: The popularity of various types of hardware systems used throughout the surveyed papers. The most popular category of robotic arms is further analyzed in

VI. Benchmarking

A. Experimental Evaluation

Most approaches implemented their grasping method in the real-world, either as an evaluation of the performance with respect to the metrics described in Section Section VI-C or presenting a demonstration that acts as a proof-of-concept of their method. Demonstrations focus on showing that their system works in the real-world without systematically evaluating their method. However, some approaches only consider evaluating their system using a simulator.

Evaluating in the real-world naturally carries more weight since the goal for robotic grasping is to be ultimately applied to real world. It should be noted that even though most works train their models using purely simulated data, they evaluate their approach in real world experiments. Commonly, the grasping system was directly transferred from simulation to real-world (e.g [37, 38, 41, 87, 88, 132]). Techniques such as domain adaptation [178] or domain randomization [179, 180] have also been used to transfer grasping approaches between simulation and real-world. Zhu et al. [97] used contrastive learning [181] to extract invariant features when images are augmented, aiming to improve the model under image sensor noise.

B. Hardware

We found that most works that study 6-DoF grasping use a robotic arm. However, a minority of works use other platforms, such as mobile platforms or humanoid robots. The robotic arms were most commonly table-mounted, and the robot would perform table-top grasping. Humanoid robots were commonly used when studying papers relating to affordances [80, 151, 152]. Surveyed works also completed research using mobile robots for grasping, however, none of the works made use of the extra DoF unique to the mobile aspect of the robot.

Most researchers used either industrial robots or robots designed for human-robot interaction. The most common robotic platform is the Franka Emika Panda. This robot has a redundant DoF which allows more freedom in joint angles when achieving a specific gripper pose.

The most commonly used gripper was a two-finger parallel jaw (51 times) with researchers using the gripper that came

Robot Hardware	Popularity	DoF
Franka Emika Panda	15	7
URXX	14	6
Kuka	14	6
Kinova	6	6 or 7
Baxter	5	7
StaubliTX60 Arm	3	6
Other	9	

TABLE VIII: Popularity of various robotic arms throughout the surveyed papers. URXX represents robotics from the UR series, including UR3, UR5, UR5e, UR10.

Gripper	Popularity	# of Fingers	DoF
Panda Gripper	12	2	1
Robotiq 2F	10	2	1
Barret Hand	7	3	5
Robotiq 3F	5	3	1
Baxter Gripper	4	2	1
Allegro	4	4	16
Kinova 2F	3	2	1
Shadow Hand	2	5	20
Kinova 3F	1	3	2
Other	34		

TABLE IX: The popularity of various grippers in the surveyed works.

with the robotic arm. Some researchers use grippers with more than two fingers, including the Barret Hand, Allegro Hand, Shadow Hand and, Kinova 3-Finger gripper. Some papers only make use of two-fingers from the Barret hand[107] or switch between two- and three-fingers[86]. Multi-fingered high-DoF hands present a different set of challenges compared to a two-finger gripper. Research on multi-fingered grippers focus on how to generate a grasp pose that considers the large amount of DoF (e.g [52, 63, 82, 90]), or grasping with affordances (e.g [121, 151, 152]). Soft grippers are another potentially interesting line of work with only a few reviewed works making use of them [45, 81, 166].

C. Performance Metrics

A diverse set of performance metrics are used among all these works. The common definitions of performance metrics related to grasping from the reviewed work are listed below. However, the exact definition of each performance metric can vary slightly between different works.

1) **Grasp Success Rate**: The percentage of successful grasps (No. of Successful Grasps / Total No. of Attempted Grasps). The post-grasp steps prior to considering a grasp ‘successful’ was not consistent across reviewed works.
2) **Completion / Clearance Rate**: The percentage of objects that are removed from the clutter (No. of Objects Grasped / Total No. of Objects in Clutter).
3) **Coverage**: The percentage of sampled ground truth grasps that are within a threshold distance of any of the generated grasps.
4) **Grasp Prediction Accuracy**: The percentage of grasps outcomes correctly predicted (No. of Successful Grasp Predictions / No. of Predictions).
5) **Computation Time**: Time required to compute grasp hypothesis generation.

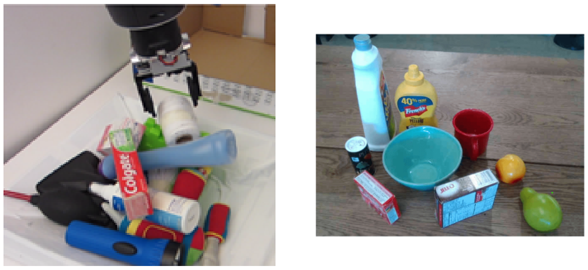


Fig. 13: Examples of cluttered scenes. We differentiate Piled Clutter (left) from Structured clutter (right). Left image is from [38] (©2017 SAGE) and Right image is from [47]. (©2020 IEEE)

6) **Precision**: The percentage of true positive grasp predictions (No. of True Positive Grasp Predictions / (No. of Selected Positive Grasps)).

D. Object Configurations

Object configurations are how the objects are arranged in the scene during training or testing. We cluster object configurations into three types.

1) **Singulated**: A single object in the scene.
2) **Piled Clutter**: Objects are packed together tightly. Objects are commonly arranged as a pile, for example Fig. 13 (Left).
3) **Structured Clutter**: Multiple objects spread out in a scene such that they are not touching, for example Fig. 13 (Right).

Most works which focus on singulated objects tend to be grasping with a high DOF hand [33, 35, 44, 48, 52, 54, 55, 67, 80, 82, 90, 100, 132, 139]. These papers tend to focus on solving a specific grasping task, rather then aiming for generalized grasping. Other works on singulated objects focus on data representation and the learning process [39, 88, 127, 131]. Singulated object grasping is also used in other contexts such as affordances [10, 15, 121] or manipulation [133]. Researchers generally do not distinguish between structured or piled clutter scenes, however, we consider these distinct scenarios, which may have different solutions.

VII. Discussion and Future Directions

This section discusses the state of the field of deep learning-based grasp synthesis and highlights recommendations for future research directions. We note the key takeaways of each subsection indicated as **Key Takeaways**.

A. Benchmarking

6-DoF grasping literature typically focuses on grasping objects from tabletop scenes where most of the synthesized grasp poses are kinematically feasible. Studying 6-DoF grasping in enclosed spaces such as shelves [182], around obstacles [50, 56], around humans [61], near reachability limits [50] or in-the-wild (e.g orchards [183]) would pose more constraints on the synthesis of the grasp configurations and require the grasp poses to not only be of high quality but also diverse, to increase the probability of finding a feasible trajectory to reach those grasp poses. As 6-DoF grasping approaches are not often tested in such challenging environments or compared directly to 4-DoF ones, the question of whether to use a 6 DoF or a 4-DoF grasping approach for a given application is not adequately answered by the current state-of-the-art.

Our review also found that few research papers provide a ready-to-use implementation of their work, making it easier for other researchers to benchmark their algorithms against. On the other hand, the ones that offer an implementation are commonly used by others as part of their evaluation. For instance, work by ten Pas et al. [38] is used by many others as a benchmark, likely because it is one of the earliest works that provide an open-source implementation wrapped in a ROS package.

Key Takeaways:

- 6-DoF grasping should be studied in more varied environments rather than just tabletop scenarios.
- Researchers should make their algorithms publicly available, ideally in a ready-to-use format (e.g. as a ROS package) to allow for informative benchmarking even on tabletop scenarios.

B. *Performance Metrics*

Bohg et al. [3] highlighted in 2014 that the grasping community has not yet embraced a consistent set of performance metrics. This observation is still valid today. This can be partly attributed to the large variety of objects, robots, end-effectors, and scenarios used in grasping research. Moreover, there is a divide between subsystem metrics and task-level metrics [184].

There are, however, some common performance metrics that the community has been using. The most popular metric is the grasp success rate, even though the exact definition varies in the literature. Some deem a grasp successful if the object is still held by the gripper after the robot returns to a configuration that is a certain height above the table [35, 88]. Other works impose additional constraints to the success definition, such as if the object is held for a certain amount of time [81] or checking if the object is still in the gripper after taking a sequence of actions intended to test the robustness of the grasp [49].

Although success rate is a useful metric, an underutilized type of performance metric is one that measures the time efficacy of a grasping approach. For instance, if the task is to remove objects from a container efficiently, is a slower robot with a 95% success rate preferred to a faster robot with only 75% success rate? The answer depends on the task, since faster grasping approach might be preferred for a task where dropping an object is not detrimental. One metric that could be used for this purpose is Mean Picks Per Hour (MPPH), which is defined as the average number of successful grasps completed in an hour. Researchers should be reminded, however, that a time-based metric such as MPPH does not only measure the computational efficiency of grasp synthesis but the system as a whole, and will be affected by the robot hardware, trajectory efficiency as well as the compute resources.

Key Takeaways:

- In addition to reporting the grasp success rate, researchers should consider reporting a more strict definition of the grasp success that tests the robustness of the grasp, similar to Lundell et al. [49].
- We suggest wider adoption of a time-based performance metric, such as the Mean Picks Per Hour (MPPH).

C. *Object Sets and Grasping Datasets*

Many papers use custom object sets, consisting of daily objects that the authors could find around the lab, making it difficult to compare different grasping approaches. Standardized object sets are very useful in enabling head-to-head comparisons of grasp synthesis algorithms. The most useful object sets for grasping research are those which cover a variety of objects and are easily accessible by the research community. Furthermore, object sets should have accurate 3D models that enable simulation studies. The most commonly used object set currently by today's grasping community is the YCB object set [154]. However, since this may change in the future, researchers should keep an eye on which object sets the community is using.

Models for grasp synthesis are trained on datasets where each data point contains the sensor data, the grasp pose executed by the robot, along with a ground truth label depending whether the grasp was successful or not. The majority of the reviewed works were trained on simulated datasets which offer large-scale data collection. However their is a lack of realism due to the physics not being modeled accurately [185]. On the other hand, there is a lack of real-world grasping datasets which offer higher-quality data but lacks the scale of simulated datasets. While we believe that a comprehensive real-world dataset could be useful for the community analogous to the commonly used Cornell Grasping dataset [104] for 4-DoF top-down grasping, most research groups typically lack the time and resources needed to create large-scale real-world grasping datasets with notable exceptions [112]. A compromise between real and simulated datasets are the hybrid datasets, for instance, Fang et al. [87] whom provided real-world point cloud data, however the grasp hypotheses are evaluated analytically rather than with actual trial-and-error.

Key Takeaways:

- Grasping researchers should adopt one of the widely used object sets. YCB object set [154] is the most commonly used one today, however, in the future, new object sets might find a wider adoption.
- The grasping community would benefit from the introduction of new object datasets that offer variety or specialization in different aspects such as object geometry, grasping difficulty [162] or deformation.
- We propose authors release the code used to generate the dataset in addition to the dataset itself. This will allow other authors to make changes to the data generation procedure.

D. *Trajectory Planning*

Motion planning is widely utilized to find collision-free trajectories to execute a chosen grasp hypothesis. However, planning for a collision-free trajectory can be too conservative for densely cluttered scene, for example in cabinets and shelves where it may not possible to grasp the object of interest without nudging the neighboring objects. It has been shown that humans make deliberate contact with the environment during manipulation rather than carefully avoiding it[186]. This strategy exploits how the environment provides physical constraints on how the object and hand move and therefore provides a funnel for uncertainty due to noise in perception and control. This principle has also been shown to work well for robots [186–193]. For example, fixtures is a widely used practice in industry for various application such as machining, assembly and inspection. This suggests that the objective of having collision-free trajectories to acquire a grasp must be relaxed.

Key Takeaway: We encourage the study of motion planning algorithms for tight spaces where it is not possible to reach the targets without nudging other objects or where contact with the environment can be leveraged for more robust grasping.

E. Sensor Modalities

Most of the reviewed papers use vision as the sole modality for perceiving the world, however, there are other sensing modalities could be very helpful for grasping. For example, tactile sensors can be used to predict if the object will remain grasped before it is actually lifted [194–196], detect object slip [197], account for uncertainty in object pos [119] and reconstruction of object geometry [44, 198]. Force/Torque sensing is another common sensor modality utilized in robotics there are also other less common modalities include robotic skins [199] and sound [200, 201].

Key Takeaway: Many of the reviewed works are vision-only sensing and considering only a single point-of-view. More research is needed to complement vision with other senses such as touch or hearing.

F. Grippers, hands and beyond

Most of the works in this survey focus on the use of simple two-fingered or three-fingered grippers, with few consider anthropomorphic hands. Simpler grippers greatly reduce the complexity of computing many-Dofs grasp hypotheses, and, most importantly, affordable commercial grippers allow researchers to setup systems for experimentally validating their results. However, this biases the research towards problems suitable for these simpler grippers and makes research on some other problems like grasping with anthropomorphic hand, dexterous, in-hand manipulation or soft-hand grasping fall behind. These directions may become important for enabling robots to go beyond pick and place tasks.

Key Takeaway: The robotics community has the potential of disrupting the manufacturing industry. Building robust and dexterous hands will result in automating many jobs currently done by humans. The opportunities are immense: new ways of designing hands, new materials, new sensors and ways of actuation, easy exchangeable fingers, redundancy. But it is not only about the hands - making sure the arm can carry the hand, that the hand is a natural extension of the arm, hand and arms in interaction- going beyond one and compensating some degrees of freedom in the hand by skillful dual-arm interaction are just some of the avenues to follow.

G. Grasping as part of a process

All day long, our hands and fingers, touch, push, pull, and enclose objects. We do this with rigid, articulated or deformable objects. We do this in the air, while objects are standing on rigid or soft surfaces or while they are moving. We do this while they are in water or covered in oil, with or without seeing - like finding a key in a pocket. Despite many decades of research and development in the area, human grasping skills are still quite superior to any of the artificial systems so far demonstrated.

In this discussion section, we have identified several areas where further contributions are needed. However, a broader view is necessary to address some if not all the manipulation scenarios mentioned above and, more importantly, consider grasping as part of an entire manipulation process. Specifically, identifying one specific 6DoF grasp pose for an object may not be sufficient to achieve success in these scenarios. One avenue may be to re-think grasp parameterizations to extend them to trajectories defined in joint and/or contact space. New ideas and techniques for online learning and recovering from failure are required. Only new ideas and techniques in these and more areas will push advances that will make robotics systems successfully perform tasks we have been promising for some time such as folding clothing, preparing food, dressing humans. Pushing for and studying grasping beyond pure grasping, is one of the most important challenge we have as a community.

In summary, we have conducted a systematic review of state-of-the-art works which use deep learning-based towards achieving 6-DoF grasping. From this review, we synthesize 10 key takeaways which we believe could enable further progress in this rapidly progressing field.

VIII. Acknowledgments

We would like to thank Wesley Chan for the helpful discussion at the ideation stages of the work. We would like to thank Lily Tran for helping with table creating and editing.

References

[1] H. Asada, “Studies on prehension and handling by robot hands with elastic fingers,” Kyoto University, Tech. Rep., 1979.

[2] A. Bicchi and V. Kumar, “Robotic grasping and contact: a review,” in *IEEE International Conference on Robotics and Automation*, vol. 1, 2000, pp. 348–353.

[3] J. Bohg, A. Morales, T. Asfour, and D. Kragic, “Data-driven grasp synthesis—a survey,” *IEEE Transactions on Robotics*, vol. 30, no. 2, p. 289–309, Apr 2014.

[4] C. Eppner, S. Höfer, R. Jonschkowski, R. Martín-Martín, A. Sieverling, V. Wall, and O. Brock, “Lessons from the amazon picking challenge: Four aspects of building robotic systems,” in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, ser. IJCAI’17. AAAI Press, 2017, p. 4831–4835.

[5] R. Deimel and O. Brock, “A novel type of compliant and underactuated robotic hand for dexterous grasping,” *The International Journal of Robotics Research*, vol. 35, no. 1-3, pp. 161–185, 2016.

[6] M. G. Catalano, G. Grioli, E. Farnioli, A. Serio, M. Bonilla, M. Garabini, C. Piazza, M. Gabiccini, and A. Bicchi, *From Soft to Adaptive Synergies: The Pisa/IIT SoftHand*. Cham: Springer International Publishing, 2016, pp. 101–125. [Online]. Available: https://doi.org/10.1007/978-3-319-26706-7_8

[7] M. Breyer, J. J. Chung, L. Ott, S. Roland, and N. Juan, “Volumetric grasping network: Real-time 6 dof grasp detection in clutter,” in *Conference on Robot Learning*, 2020.

[8] D. Morrison, P. Corke, and J. Leitner, "Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach," in *Proc. of Robotics: Science and Systems (RSS)*, 2018.

[9] H. Kang and C. Chen, "Fruit detection and segmentation for apple harvesting using visual sensor in orchards," *Sensors*, vol. 19, p. 4599, 10 2019.

[10] A. Murali, W. Liu, K. Marino, S. Chernova, and A. Gupta, "Same object, different grasps: Data and semantic knowledge for task-oriented grasping," in *Conference on Robot Learning*, 2020.

[11] K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber, "A survey on learning-based robotic grasping," *Current Robotics Reports*, vol. 1, no. 4, pp. 239–249, 2020.

[12] H. Duan, P. Wang, Y. Huang, G. Xu, W. Wei, and X. Shen, "Robotics dexterous grasping: The methods based on point cloud and deep learning," *Frontiers in Neurorobotics*, vol. 15, p. 73, 2021.

[13] O. Kroemer, S. Niekum, and G. D. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," *J. Mach. Learn. Res.*, vol. 22, pp. 30:1–30:82, 2021.

[14] J. J. Gibson, "The theory of affordances," in *Perceiving, acting, and knowing: toward an ecological psychology*, J. B. Robert E Shaw, Ed. Hillsdale, N.J. : Lawrence Erlbaum Associates, 1977, pp. pp.67–82. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00692033

[15] P. Ardón, È. Pairet, R. Petrick, S. Ramamoorthy, and K. Lohan, "Learning grasp affordance reasoning through semantic relations," *IEEE Robotics and Automation Letters*, vol. PP, pp. 1–1, 08 2019.

[16] A. Morales, T. Asfour, P. Azad, S. Knoop, and R. R. Dillmann, "Integrated grasp planning and visual object localization for a humanoid robot with five-fingered hands," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Beijing, China: Ieee, 2006, pp. 5663–5668.

[17] V.-D. Nguyen, "Constructing force-closure grasps in 3d," in *IEEE International Conference on Robotics and Automation*, vol. 4, 1987, pp. 240–245.

[18] K. Mirza and D. Grin, "General formulation for force distribution in power grasp," in *IEEE International Conference on Robotics and Automation*, 1994, pp. 880–887 vol.1.

[19] I.-M. Chen and J. Burdick, "Finding antipodal point grasps on irregularly shaped objects," *IEEE Transactions on Robotics and Automation*, vol. 9, no. 4, pp. 507–512, 1993.

[20] D. Prattichizzo and J. C. Trinkle, "Grasping," in *Springer handbook of robotics*. Springer, 2016, pp. 955–988.

[21] R. Murray, Z. Li, S. Sastry, and S. Sastry, *A Mathematical Introduction to Robotic Manipulation*. Taylor & Francis, 1994.

[22] A. Miller and P. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robotics Automation Magazine*, vol. 11, no. 4, pp. 110–122, 2004.

[23] S. El-Khoury, A. Sahbani, and V. Perdereau, "Learning the natural grasping component of an unknown object," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007, pp. 2957–2962.

[24] F. Kyota, T. Watabe, S. Saito, and M. Nakajima, "Detection and evaluation of grasping positions for autonomous agents," in *International Conference on Cyberworlds*, 2005.

[25] C. Michel, V. Perdereau, and M. Drouin, "An approach to extract natural grasping axes with a real 3d vision system," in *IEEE International Symposium on Industrial Electronics*, vol. 4, 2006, pp. 3130–3135.

[26] I. Kamon, T. Flash, and S. Edelman, "Learning to grasp using visual information," in *Proceedings of IEEE International Conference on Robotics and Automation*, vol. 3, 1996.

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

[28] A. Saxena, J. Driemeyer, J. Kearns, C. Osondu, and A. Y. Ng, "Learning to grasp novel objects using vision," in *Experimental Robotics*. Springer, 2008, pp. 33–42.

[29] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, 2008.

[30] A. Saxena, L. Wong, and A. Ng, "Learning grasp strategies with partial shape information." in *Proceedings of the National Conference on Artificial Intelligence*, vol. 3, 2008, pp. 1491–1494.

[31] A. Morales, E. Chinellato, P. Sanz, A. P. del Pobil, and A. Fagg, "Learning to predict grasp reliability for a multifinger robot hand by using visual features," 2004.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[33] U. R. Aktas, C. Zhao, M. Kopicki, A. Leonardis, and J. L. Wyatt, "Deep dexterous grasping of novel objects from a single view," *arXiv preprint arXiv:1908.04293*, 2019.

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

[35] J. Varley, J. Weisz, J. Weiss, and P. Allen, "Generating multi-fingered robotic grasps via deep learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 4415–4420.

[36] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 4304–4311.

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

[38] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.

[39] Y. Zhou and K. Hauser, "6dof grasp planning by optimizing a deep learning scoring function," in *Robotics: Science and Systems Workshop on Revisiting Contact-Turning a Problem into a Solution*, 2017.

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

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

[42] H. Liang, X. Ma, S. Li, M. Gorner, S. Tang, B. Fang, F. Sun, and J. Zhang, "Pointnetgpd: Detecting grasp configurations from point sets," *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[43] Q. Lu, K. Chenna, B. Sundaralingam, and T. Hermans, "Planning multi-fingered grasps as probabilistic inference in a learned deep network," in *Robotics Research*. Springer, 2019, pp. 455–472.

[44] S. Ottenhaus, D. Renninghoff, R. Grimm, F. Ferreira, and T. Asfour, "Visuo-haptic grasping of unknown objects based on gaussian process implicit surfaces and deep learning," in *IEEE-RAS International Conference on Humanoid Robots*, 2019, pp. 402–409.

[45] J. Gonçalves and P. Lima, "Grasp planning with incomplete knowledge about the object to be grasped," in *IEEE International Conference on Autonomous Robot Systems and Competitions*, 2019, pp. 1–6.

[46] M. A. c. Riedlinger, M. Voelk, K. Kleeberger, M. U. Khalid, and R. Bormann, "Model-free grasp learning framework based on physical simulation," in *International Symposium on Robotics*, 2020, pp. 1–8.

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

[48] M. Van der Merwe, Q. Lu, B. Sundaralingam, M. Matak, and T. Hermans, "Learning continuous 3d reconstructions for geometrically aware grasping," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 11 516–11 522.

[49] J. Lundell, F. Verdoja, and V. Kyrki, "Beyond top-grasps through scene completion," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 545–551.

[50] X. Lou, Y. Yang, and C. Choi, "Learning to generate 6-dof grasp poses with reachability awareness," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1532–1538.

[51] Y. Choi, H. Kee, K. Lee, J. Choy, J. Min, S. Lee, and S. Oh, "Hierarchical 6-dof grasping with approaching direction selection," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1553–1559.

[52] Q. Lu, M. Van der Merwe, B. Sundaralingam, and T. Hermans, "Multifingered grasp planning via inference in deep neural networks: Outperforming sampling by learning differentiable models," *IEEE Robotics & Automation Magazine*, vol. 27, no. 2, pp. 55–65, 2020.

[53] H. Schaub and A. Schöttl, "6-dof grasp detection for unknown objects," in *International Conference on Advanced Computer Information Technologies (ACIT)*, 2020, pp. 400–403.

[54] M. Kokic, D. Kragic, and J. Bohg, "Learning task-oriented grasping from human activity datasets," *IEEE Robotics and Automation Letters*, vol. PP, pp. 1–1, 02 2020.

[55] J. Lundell, E. Corona, T. N. Le, F. Verdoja, P. Weinzaepfel, G. Rogez, F. Moreno-Noguer, and V. Kyrki, "Multi-fingan: Generative coarse-to-fine sampling of multi-finger grasps," *arXiv preprint arXiv:2012.09696*,

2020.

[56] X. Lou, Y. Yang, and C. Choi, “Collision-aware target-driven object grasping in constrained environments,” *CoRR*, vol. abs/2104.00776, 2021.

[57] J. Lundell, F. Verdoja, and V. Kyrki, “Ddgc: Generative deep dexterous grasping in clutter,” 2021.

[58] G. Peng, Z. Ren, H. Wang, X. Li, and M. O. Khyam, “A self-supervised learning-based 6-dof grasp planning method for manipulator,” *IEEE Transactions on Automation Science and Engineering*, 2021.

[59] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, “Synergies between affordance and geometry: 6-dof grasp detection via implicit representations,” 2021.

[60] H. Kasaei and M. Kasaei, “Mvgrasp: Real-time multi-view 3d object grasping in highly cluttered environments,” 2021.

[61] X. Wang, S. Nisar, and F. Matsuno, “Robust grasp detection with incomplete point cloud and complex background,” *Advanced Robotics*, vol. 0, no. 0, pp. 1–16, 2021.

[62] M. R. Munoz, “Grasping in 6dof: An orthographic approach to generalized grasp affordance predictions,” *https://fse.studenttheses.ub.rug.nl/*, 2021.

[63] M. Corsaro, S. Tellex, and G. Konidaris, “Learning to detect multi-modal grasps for dexterous grasping in dense clutter,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 4647–4653.

[64] Z. Ren, G. Peng, J. Yang, and H. Wang, “6-dof grasp planning of manipulator combined with self-supervised learning,” in *Chinese Control and Decision Conference (CCDC)*, 2021, pp. 3026–3032.

[65] B. Wen, W. Lian, K. Bekris, and S. Schaal, “Catgrasp: Learning category-level task-relevant grasping in clutter from simulation,” in *Arxiv*, 2021.

[66] 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 *IEEE International Conference on robotics and automation (ICRA)*, 2018, pp. 5620–5627.

[67] Q. Lu, M. Van der Merwe, and T. Hermans, “Multi-fingered active grasp learning,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 8415–8422.

[68] A. Pas and R. Platt, *Using Geometry to Detect Grasp Poses in 3D Point Clouds*, 01 2018, pp. 307–324.

[69] G. Smith, E. Lee, K. Goldberg, K. F. Böhringer, and J. J. Craig, “Computing parallel-jaw grips,” *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No.99CH36288C)*, vol. 3, pp. 1897–1903 vol.3, 1999.

[70] S. Agrawal and N. Goyal, “Analysis of thompson sampling for the multi-armed bandit problem,” in *Conference on learning theory*. JMLR Workshop and Conference Proceedings, 2012, pp. 39–1.

[71] N. Vahrenkamp, M. Kröhnert, S. Ulbrich, T. Asfour, G. Metta, R. Dillmann, and G. Sandini, *Simox: A Robotics Toolbox for Simulation, Motion and Grasp Planning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 585–594.

[72] M. Ciocarlie and P. Allen, “Hand posture subspaces for dexterous robotic grasping,” *I. J. Robotic Res.*, vol. 28, pp. 851–867, 06 2009.

[73] M. S. Kopicki, D. Belter, and J. L. Wyatt, “Learning better generative models for dexterous, single-view grasping of novel objects,” *The International Journal of Robotics Research*, vol. 38, no. 10-11, pp. 1246–1267, 2019. [Online]. Available: https://doi.org/10.1177/0278364919865338

[74] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” 2014.

[75] C. M. Bishop, “Mixture density networks,” Aston University, Tech. Rep., 1994.

[76] H. Schaub, A. Schöttl, and M. Hoh, “6-dof grasp detection for unknown objects using surface reconstruction,” in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 2021, pp. 1–6.

[77] V. Satish, J. Mahler, and K. Goldberg, “On-policy dataset synthesis for learning robot grasping policies using fully convolutional deep networks,” *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1357–1364, 2019.

[78] C. Ferrari and J. Canny, “Planning optimal grasps,” in *Proceedings 1992 IEEE International Conference on Robotics and Automation*, 1992, pp. 2290–2295 vol.3.

[79] M. Veres, M. Moussa, and G. W. Taylor, “Modeling grasp motor imagery through deep conditional generative models,” *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 757–764, 2017.

[80] P. Schmidt, N. Vahrenkamp, M. Wächter, and T. Asfour, “Grasping of unknown objects using deep convolutional neural networks based on depth images,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6831–6838.

[81] C. Choi, W. Schwarting, J. DelPreto, and D. Rus, “Learning object grasping for soft robot hands,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2370–2377, 2018.

[82] M. Liu, Z. Pan, K. Xu, K. Ganguly, and D. Manocha, “Generating grasp poses for a high-dof gripper using neural networks,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1518–1525.

[83] 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*. PMLR, 2020, pp. 53–65.

[84] D. Yang, T. Tosun, B. Eisner, V. Isler, and D. Lee, “Robotic grasping through combined image-based grasp proposal and 3d reconstruction,” 2020.

[85] K.-Y. Jeng, Y.-C. Liu, Z. Y. Liu, J.-W. Wang, Y.-L. Chang, H.-T. Su, and W. H. Hsu, “Gdn: A coarse-to-fine (c2f) representation for end-to-end 6-dof grasp detection,” 2020.

[86] C.-H. Wang and P.-C. Lin, “Q-pointnet: Intelligent stacked-objects grasping using a rgbd sensor and a dexterous hand,” in *IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE Press, 2020, p. 601–606.

[87] H.-S. Fang, C. Wang, M. Gou, and C. Lu, “Graspnet-1billion: a large-scale benchmark for general object grasping,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[88] C. Wu, J. Chen, Q. Cao, J. Zhang, Y. Tai, L. Sun, and K. Jia, “Grasp proposal networks: An end-to-end solution for visual learning of robotic grasps,” in *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[89] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3619–3625.

[90] M. Liu, Z. Pan, K. Xu, K. Ganguly, and D. Manocha, “Deep differentiable grasp planner for high-dof grippers,” *CoRR*, vol. abs/2002.01530, 2020.

[91] L. Shao, F. Ferreira, M. Jorda, V. Nambiar, J. Luo, E. Solowjow, J. A. Ojea, O. Khatib, and J. Bohg, “Unigrasp: Learning a unified model to grasp with multifingered robotic hands,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2286–2293, 2020.

[92] P. Ni, W. Zhang, X. Zhu, and Q. Cao, “Learning an end-to-end spatial grasp generation and refinement algorithm from simulation,” *Machine Vision and Applications*, vol. 32, no. 1, pp. 1–12, 2021.

[93] C. Wang, H.-S. Fang, M. Gou, H. Fang, J. Gao, and C. Lu, “Graspness discovery in clutters for fast and accurate grasp detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 15 964–15 973.

[94] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, “Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes,” 2021.

[95] 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,” *arXiv preprint arXiv:2002.12647*, 2020.

[96] M. Gou, H. Fang, Z. Zhu, S. Xu, C. Wang, and C. Lu, “RGB matters: Learning 7-dof grasp poses on monocular RGBD images,” *CoRR*, vol. abs/2103.02184, 2021.

[97] X. Zhu, L. Sun, Y. Fan, and M. Tomizuka, “6-dof contrastive grasp proposal network,” *CoRR*, vol. abs/2103.15995, 2021.

[98] W. Wei, Y. Luo, F. Li, G. Xu, J. Zhong, W. Li, and P. Wang, “Gpr: Grasp pose refinement network for cluttered scenes,” 2021.

[99] Y. Li, T. Kong, R. Chu, Y. Li, P. Wang, and L. Li, “Simultaneous semantic and collision learning for 6-dof grasp pose estimation,” 2021.

[100] S. Li, Z. Li, K. Han, X. Li, Y. Xiong, and Z. Xie, “An end-to-end spatial grasp prediction model for humanoid multi-fingered hand using deep network,” in *2021 6th International Conference on Control, Robotics and Cybernetics (CRC)*, 2021, pp. 130–136.

[101] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *CVPR*, 06 2016, pp. 779–788.

[102] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

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

[104] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *International Journal of Robotics Research*, vol. 34, 01 2013.

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

[106] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 769–776.

[107] B. Wu, I. Akinola, A. Gupta, F. Xu, J. Varley, D. Watkins-Valls, and P. K. Allen, "Generative attention learning: a "general" framework for high-performance multi-fingered grasping in clutter," *Autonomous Robots*, pp. 1–20, 2020.

[108] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *NIPS*, 2015.

[109] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5105–5114.

[110] Y. Li, "Deep reinforcement learning: An overview," 2017. [Online]. Available: https://arxiv.org/abs/1701.07274

[111] M. Mohammed, L. Kwek, and S. C. Chua, "Review of deep reinforcement learning-based object grasping: Techniques, open challenges, and recommendations," *IEEE Access*, vol. 8, pp. 178 450–178 481, 10 2020.

[112] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 421–436, 2018.

[113] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," in *Conference on Robot Learning (CoRL)*, 06 2018.

[114] D. Quillen, E. Jang, O. Nachum, C. Finn, J. Ibarz, and S. Levine, "Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6284–6291.

[115] M. Gualtieri and R. Platt, "Learning 6-dof grasping and pick-place using attention focus," in *Conference on Robot Learning*. PMLR, 2018, pp. 477–486.

[116] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[117] B. Wu, I. Akinola, and P. K. Allen, "Pixel-attentive policy gradient for multi-fingered grasping in cluttered scenes," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1789–1796.

[118] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.

[119] H. Merzic, M. Bogdanovic, D. Kappler, L. Righetti, and J. Bohg, "Leveraging contact forces for learning to grasp," in *2019 IEEE International Conference on Robotics and Automation*, 2019.

[120] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," 2015. [Online]. Available: https://arxiv.org/abs/1502.05477

[121] P. Mandikal and K. Grauman, "Learning dexterous grasping with object-centric visual affordances," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[122] S. Song, A. Zeng, J. Lee, and T. Funkhouser, "Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations," *IEEE Robotics and Automation Letters*, vol. PP, pp. 1–1, 06 2020.

[123] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[124] L. Berscheid, C. Friedrich, and T. Kröger, "Robot learning of 6 dof grasping using model-based adaptive primitives," *CoRR*, vol. abs/2103.12810, 2021.

[125] D. Kawakami, R. Ishikawa, M. Roxas, Y. Sato, and T. Oishi, "Learning 6dof grasping using reward-consistent demonstration," 2021.

[126] B. Tang, M. Corsaro, G. Konidaris, S. Nikolaidis, and S. Tellex, "Learning collaborative pushing and grasping policies in dense clutter," 2021.

[127] L. Wang, Y. Xiang, and D. Fox, "Goal-auxiliary actor-critic for 6d robotic grasping with point clouds," 2021.

[128] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[129] X. Chen, Z. Ye, J. Sun, Y. Fan, F. Hu, C. Wang, and C. Lu, "Transferable active grasping and real embodied dataset," 2020.

[130] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "Contactdb: Analyzing and predicting grasp contact via thermal imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8709–8719.

[131] T. Patten, K. Park, and M. Vincze, "Dgcm-net: Dense geometrical correspondence matching network for incremental experience-based robotic grasping," *Frontiers in Robotics and AI*, vol. 7, p. 120, 2020.

[132] J. Varley, C. DeChant, A. Richardson, J. Ruales, and P. Allen, "Shape completion enabled robotic grasping," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 2442–2447.

[133] W. Gao and R. Tedrake, "kpam-sc: Generalizable manipulation planning using keypoint affordance and shape completion," 2019.

[134] X. Zhang, Z. Zhang, C. Zhang, J. B. Tenenbaum, W. T. Freeman, and J. Wu, "Learning to reconstruct shapes from unseen classes," 2018.

[135] M. Kiatos, S. Malassiotis, and I. Sarantopoulos, "A geometric approach for grasping unknown objects with multifingered hands," *IEEE Transactions on Robotics*, pp. 1–12, 2020.

[136] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, p. 307–392, 2019.

[137] N. Chavan-Dafle, S. Popovych, S. Agrawal, D. D. Lee, and V. Isler, "Object shell reconstruction: Camera-centric object representation for robotic grasping," *arXiv preprint arXiv:2109.06837*, 2021.

[138] M. Gualtieri and R. Platt, "Robotic pick-and-place with uncertain object instance segmentation and shape completion," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1753–1760, 2021.

[139] J. Lundell, F. Verdoja, and V. Kyrki, "Robust grasp planning over uncertain shape completions," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1526–1532.

[140] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*, 2016, pp. 1050–1059.

[141] P. Vafaeikia, K. Namdar, and F. Khalvati, "A brief review of deep multi-task learning and auxiliary task learning," *arXiv preprint arXiv:2007.01126*, 2020.

[142] Y. Avigal, S. Paradis, and H. Zhang, "6-dof grasp planning using fast 3d reconstruction and grasp quality cnn," *arXiv preprint arXiv:2009.08618*, 2020.

[143] 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," in *Robotics: Science and Systems (RSS)*, 2017.

[144] M. Björkman, Y. Bekiroglu, V. Hogman, and D. Kragic, "Enhancing visual perception of shape through tactile glances," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3180–3186, 2013.

[145] Y. Shiraki, K. Nagata, N. Yamanobe, A. Nakamura, K. Harada, D. Sato, and D. N. Nenchev, "Modeling of everyday objects for semantic grasp," in *IEEE International Symposium on Robot and Human Interactive Communication*, 2014, pp. 750–755.

[146] A. Miller, S. Knoop, H. Christensen, and P. Allen, "Automatic grasp planning using shape primitives," in *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*, vol. 2, 2003, pp. 1824–1829 vol.2.

[147] T. Torii and M. Hashimoto, "Model-less estimation method for robot grasping parameters using 3d shape primitive approximation," in *IEEE International Conference on Automation Science and Engineering (CASE)*, 2018, pp. 580–585.

[148] P. Ardón, É. Pairet, K. S. Lohan, S. Ramamoorthy, and R. Petrick, "Affordances in robotic tasks–a survey," *arXiv preprint arXiv:2004.07400*, 2020.

[149] Y. Li, L. Schomaker, and S. H. Kasaei, "Learning to grasp 3d objects using deep residual u-nets," in *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2020, pp. 781–787.

[150] T.-T. Do, A. Nguyen, and I. Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 5882–5889.

[151] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Detecting object affordances with convolutional neural networks," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 2765–2770.

[152] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Object-based affordances detection with convolutional neural networks and dense conditional random fields," in *IEEE/RSJ International*

Conference on Intelligent Robots and Systems (IROS), 2017, pp. 5908–5915.

[153] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, "kpam: Keypoint affordances for category-level robotic manipulation," *arXiv preprint arXiv:1903.06684*, 2019.

[154] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *International Conference on Advanced Robotics (ICAR)*, 2015, pp. 510–517.

[155] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel, "BigBIRD: A large-scale 3D database of object instances," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 509–516.

[156] A. Kasper, Z. Xue, and R. Dillmann, "The kit object models database: An object model database for object recognition, localization and manipulation in service robotics," *The International Journal of Robotics Research*, vol. 31, no. 8, pp. 927–934, 2012.

[157] A. X. Chang *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[158] W. Wohlkinger, A. Aldoma Buchaca, R. Rusu, and M. Vincze, "3DNet: Large-Scale Object Class Recognition from CAD Models," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2012.

[159] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The princeton shape benchmark," in *Shape Modeling Applications*, 2004, pp. 167–388.

[160] Zhirong Wu, S. Song, A. Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and J. Xiao, "3D ShapeNets: A deep representation for volumetric shapes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1912–1920.

[161] Y. Xiang, W. Kim, W. Chen, J. Ji, C. Choy, H. Su, R. Mottaghi, L. Guibas, and S. Savarese, "Objectnet3d: A large scale database for 3d object recognition," in *European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 160–176.

[162] D. Morrison, P. Corke, and J. Leitner, "Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4368–4375, 2020.

[163] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4243–4250.

[164] D. Wang, D. Tseng, P. Li, Y. Jiang, M. Guo, M. Danielczuk, J. Mahler, J. Ichnowski, and K. Goldberg, "Adversarial grasp objects," in *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, 2019, pp. 241–248.

[165] C. Eppner, A. Mousavian, and D. Fox, "Acronym: A large-scale grasp dataset based on simulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6222–6227.

[166] D. S. Diaz Cortes, G. Hwang, and K.-U. Kyung, "Imitation learning based soft robotic grasping control without precise estimation of target posture," in *International Conference on Soft Robotics*, 2021, pp. 149–154.

[167] V. R. Osorio, R. Iyengar, X. Yao, P. Bhattachan, A. Ragobar, N. Dey, and B. Tripp, "37,000 human-planned robotic grasps with six degrees of freedom," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3346–3351, 2020.

[168] L. Wang, Y. Xiang, and D. Fox, "Manipulation trajectory optimization with online grasp synthesis and selection," *arXiv preprint arXiv:1911.10280*, 2019.

[169] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "Grab: A dataset of whole-body human grasping of objects," in *Computer Vision – ECCV 2020*, vol. LNCS 12355. Cham: Springer International Publishing, Aug. 2020, pp. 581–600. [Online]. Available: https://grab.is.tue.mpg.de

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

[171] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[172] M. Bajracharya *et al.*, "A mobile manipulation system for one-shot teaching of complex tasks in homes," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 11 039–11 045.

[173] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[174] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[175] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[176] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[177] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 922–928.

[178] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," *CoRR*, vol. abs/1709.07857, 2017. [Online]. Available: http://arxiv.org/abs/1709.07857

[179] J. Tobin, L. Biewald, R. Duan, M. Andrychowicz, A. Handa, V. Kumar, B. McGrew, J. Schneider, P. Welinder, W. Zaremba, and P. Abbeel, "Domain randomization and generative models for robotic grasping," 2017. [Online]. Available: https://arxiv.org/abs/1710.06425

[180] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis, "Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks," 2018. [Online]. Available: https://arxiv.org/abs/1812.07252

[181] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," 2020. [Online]. Available: https://arxiv.org/abs/2004.11362

[182] M. Costanzo, G. De Maria, G. Lettera, and C. Natale, "Can robots refill a supermarket shelf?: Motion planning and grasp control," *IEEE Robotics Automation Magazine*, vol. 28, no. 2, pp. 61–73, 2021.

[183] X. Wang, H. Kang, H. Zhou, W. Au, and C. Chen, "Geometry-aware fruit grasping estimation for robotic harvesting in apple orchards," *Computers and Electronics in Agriculture*, vol. 193, p. 106716, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169922000333

[184] J. Leitner *et al.*, "The acrv picking benchmark: A robotic shelf picking benchmark to foster reproducible research," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4705–4712, 2017.

[185] C. Rubert, D. Kappler, J. Bohg, and A. Morales, "Predicting grasp success in the real world - a study of quality metrics and human assessment," *Robotics and Autonomous Systems*, vol. 121, p. 103274, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889019300247

[186] R. Deimel, C. Eppner, J. Álvarez-Ruiz, M. Maertens, and O. Brock, "Exploitation of environmental constraints in human and robotic grasping," in *Robotics Research*. Springer, 2016, pp. 393–409.

[187] M. Kazemi, J.-S. Valois, J. A. Bagnell, and N. Pollard, "Robust object grasping using force compliant motion primitives," 2021.

[188] L. Righetti, M. Kalakrishnan, P. Pastor, J. Binney, J. Kelly, R. C. Voorhies, G. S. Sukhatme, and S. Schaal, "An autonomous manipulation system based on force control and optimization," *Autonomous Robots*, vol. 36, no. 1-2, pp. 11–30, 2014.

[189] N. Hudson, T. Howard, J. Ma, A. Jain, M. Bajracharya, S. Myint, C. Kuo, L. Matthies, P. Backes, P. Hebert *et al.*, "End-to-end dexterous manipulation with deliberate interactive estimation," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 2371–2378.

[190] M. Toussaint, N. Ratliff, J. Bohg, L. Righetti, P. Englert, and S. Schaal, "Dual execution of optimized contact interaction trajectories," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 47–54.

[191] N. C. Dafle, A. Rodriguez, R. Paolini, B. Tang, S. S. Srinivasa, M. Erdmann, M. T. Mason, I. Lundberg, H. Staab, and T. Fuhlbrigge, "Extrinsic dexterity: In-hand manipulation with external forces," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 1578–1585.

[192] N. Chavan-Dafle and A. Rodriguez, "Prehensile pushing: In-hand manipulation with push-primitives," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 6215–6222.

[193] L. Shao, T. Migimatsu, and J. Bohg, "Learning to scaffold the development of robotic manipulation skills," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp.

5671–5677.

[194] 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.

[195] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, 2018.

[196] L. Chumbley, M. Gu, R. Newbury, J. Leitner, and A. Cosgun, "Integrating high-resolution tactile sensing into grasp stability prediction," in *Conference on Robots and Vision*, 2022.

[197] W. Chen, H. Khamis, I. Birznieks, N. F. Lepora, and S. J. Redmond, "Tactile sensors for friction estimation and incipient slip detection—toward dexterous robotic manipulation: A review," *IEEE Sensors Journal*, vol. 18, no. 22, pp. 9049–9064, 2018.

[198] C. de Farias, N. Marturi, R. Stolkin, and Y. Bekiroglu, "Simultaneous tactile exploration and grasp refinement for unknown objects," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3349–3356, 2021.

[199] D. Hughes, J. Lammie, and N. Correll, "A robotic skin for collision avoidance and affective touch recognition," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1386–1393, 2018.

[200] G. Zöller, V. Wall, and O. Brock, "Active acoustic contact sensing for soft pneumatic actuators," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 7966–7972.

[201] D. Gandhi, A. Gupta, and L. Pinto, "Swoosh! rattle! thump!–actions that sound," *arXiv preprint arXiv:2007.01851*, 2020.

[202] M. Shishegar, D. Kerr, and J. Blake, "A systematic review of research into how robotic technology can help older people," *Smart Health*, vol. 7, 2018.

IX. Appendix

A. Appendix A: Methodology

We conducted a Systematic Literature Review to assess the state-of-the-art of 6-DoF robotic grasping utilizing deep learning approaches. Our review, methodology adapted from [202], searched through six different scholarly libraries – IEEE, Springer, ScienceDirect, SpringerLink, arXiv and Taylor & Francis – with the following search terms and criteria.

B. Search Terms

The following four search terms, combined with the OR function:

- 6-DoF Grasping and
- Grasping in metadata, if metadata search was available
- Grasping AND Point Cloud AND Deep Learning
- Shape Completion AND 6-DoF Grasping AND Learning
- Affordances AND Grasping AND 6 DoF

C. Inclusion/Exclusion Criteria

Only publications that met the following criteria are included in this review:

(a) Paper considered grasping from table scenario,
(b) All 6-DoF were used for the grasp pose,
(c) Deep Learning methods were applied in an aspect of the work,
(d) Published after Jan 1, 2012, (the year Alexnet [32] was published) and
(e) Written in English.

Our search returned a total of xxx (85?) papers that matched these criteria and were therefore included in this review (see Appendix A for full list).

D. Data Extraction/Analysis

Data was extracted from the included papers and verified by at least two of the authors.

E. Impact Score

We define an Impact score for each paper, inspired by [202]. The impact score, the sum of three terms, while not fully reflective of the objective impact of the paper, provides a strong indication of a methods impact and quality.

1) Citations Per Year, normalized between [0, 1].
2) Impact Factor: the journal/conference impact factor, normalized between [0,1]. If the paper was not published, but only on pre-print archives (such as arXiV) they were given an impact factor of zero. This was to favor peer-reviewed articles.
3) Real World Testing:
   a) 0 - Simulation Only: The work includes no real world components
   b) 0.5 - Real-World Demonstration Only: They show a proof-of-concept in the real-world without a methodical testing method.
   c) 1 - Real World Testing: Testing the on a real robot employing empirical methods.

As grasping is a robotics task being able to apply the method in the real-world is important. Validating the work using empirical methods further demonstrates the strength of the research.

Rhys Newbury is a PhD student at Monash University, Australia. He holds a B. Eng and B. Sc (Computer Science) from Monash University. His research interests focus around manipulation and human-robot interaction. He has a focus on systems view to problems and how to apply robots to real-world issues.

Morris Gu is a PhD student at Monash University, Australia. He holds a Bachelor of Engineering from Monash University. His research interests are in explainability and augmented reality within human-robot interaction.

Lachlan Chumbley is an undergraduate student at Monash University, Australia. His research interests include robotic grasping and manipulation, particularly though the use of tactile information.

Arsalan Mousavian received the B.Sc. degree from the Iran University of Science and Technology, Tehran, Iran, in 2010, the M.Sc. degree from the University of Tehran, Tehran, in 2013, both in AI and robotics, and the Ph.D. degree in computer science from George Mason University, Fairfax, VA, USA, in 2018. He is currently a Senior Research Scientist with NVIDIA, Seattle, WA, USA. His research interests include 3-D perception methods that help robots accomplish robot manipulation tasks in the real world.

Clemens Eppner is a Research Scientist in the Seattle Robotics Lab at NVIDIA Research. He is interested in the problem space of grasping and manipulation, including aspects of planning, control, and perception. Before joining NVIDIA, he received his Ph.D. at the Robotics and Biology Lab at TU Berlin. While studying at the University of Freiburg, he wrote his Master's thesis at the Autonomous Intelligent Systems.

Jürgen Leitner is co-founder at LYRO Robotics, Australia. He has designed, developed, and deployed robots fore more than 15 years. His interest lies in building intelligent robots that can safely and reliably interact with the physical world. He holds a PhD (2014), a MSc (Space Science, 2009), a MSc(Tech) (Space Robotics, 2009) and a BSc (Software Engineering, 2007).

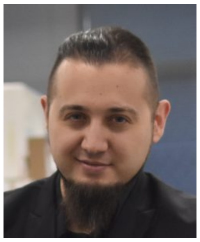
Jeannette Bohg is an Assistant Professor of Computer Science at Stanford University. She was a group leader at the Autonomous Motion Department (AMD) of the MPI for Intelligent Systems until September 2017. Before joining AMD in January 2012, Jeannette Bohg was a PhD student at the Division of Robotics, Perception and Learning (RPL) at KTH in Stockholm. In her thesis, she proposed novel methods towards multi-modal scene understanding for robotic grasping. Her research focuses on perception and learning for autonomous robotic manipulation and grasping.

Antonio Morales is Associate Professor at the Department of Computer Engineering and Science in the Universitat Jaume I of Castelló, Spain. He received his PhD in Computer Science Engineering from Universitat Jaume I in January 2004. Currently, he is director of the Degree Program on Robotics Intelligence and vice-dean of the School on Technology and Experimental Sciences of the Universitat Jaume I He is a leading researcher at the Robotic Intelligence Laboratory at Universitat Jaume I and his research interests are focused on reactive robot grasping and manipulation. He has been a Principal Investigator on several European and national research projects. He has supervised 4 PhD students.

Tamim Asfour is Professor at the Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT). He received his diploma degree in Electrical Engineering and his PhD in Computer Science from the University of Karlsruhe (TH). He is developer and leader of the development team of the ARMAR humanoid robot family. He is European Chair of the IEEE RAS Technical Committee on Humanoid Robots and member the Executive Board of the German Robotics Association (DGR: Deutsche Gesellschaft für Robotik)

Danica Kragic is a Professor at the School of Computer Science and Communication at KTH in Stockholm. She received MSc in Mechanical Engineering from the Technical University of Rijeka, Croatia in 1995 and PhD in Computer Science from KTH in 2001. Danica received the 2007 IEEE Robotics and Automation Society Early Academic Career Award. She is a member of the Swedish Royal Academy of Sciences and Swedish Young Academy. She has chaired the IEEE RAS Technical Committee on Computer and Robot Vision and from 2009 serves as an IEEE RAS AdCom member.

Dieter Fox (Fellow, IEEE) received the Ph.D. degree in computer science from the University of Bonn, Bonn, Germany, in 1998. He is currently a Professor with the Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA, where he heads the UW Robotics and State Estimation Lab. He is also Senior Director of Robotics Research with NVIDIA.He has authored or coauthored more than 200 technical papers. He is a co-author of the textbook entitled Probabilistic Robotics (Cambridge, MA, USA: MIT Press, 2005). His research interests include robotics and artificial intelligence, with a focus on state estimation and perception applied to problems such as mapping, object detection and tracking, manipulation, and activity recognition.



Akansel Cosgun received the Ph.D. degree in robotics from the Georgia Institute of Technology, Atlanta, GA, USA, in 2016. Since 2018, he has been a Research Fellow with Monash University, Clayton, VIC, Australia. He conducts research in robotics, human–robot interaction, and robot learning. From 2019 to 2020, he was the Team Lead for Vision-Based Manipulation with the Australian Centre for Robotic Vision. He has previously worked with Honda Research, Toyota Infotechnology Center, Microsoft Research, and Savioke, a robotics start-up. His research interests include mobile robots, robotic arms, and self-driving cars with an emphasis on a systems view to problems.