BIG-Net: Deep Learning for Grasping with a Bio-Inspired Soft Gripper

Hui Zhang, *Member, IEEE*, Yanming Wu, Eric Demeester, Karel Kellens

***Abstract*—In this letter, a grasping neural network for a bio-inspired gripper (BIG-Net) trained on a synthetic dataset is proposed for the picking of novel objects. The grasp feasibility is evaluated by tracking the deformation of the soft gripping pad and three types of gripping forces during simulation. Over 420 K grasp scenes with 4.3 B grasps have been synthesized with stacked objects to train the neural network, instead of isolated objects in many existing methods. The BIG-Net takes in a depth image and provides pixel-wise grasp parameters for a grasp scene. Various experiments in both simulation and real world indicate that the BIG-Net grasping method outperforms the traditional and state-of-the-art methods. It achieves the average grasp success rates of 94% for the random picking of household items in clutter and 86% for adversarial items at real-time speeds (25 ms).**

***Index Terms*—Contact modeling, deep learning in grasping and manipulation, grasp simulation, perception for grasping and manipulation.**

I. Introduction

AUTOMATIC grasping of novel objects is an essential skill for robots and remains challenging due to robot's and gripper's limitations, and environmental uncertainties. Bio-inspired soft grippers, such as an octopus gripper and a gecko gripper, have more advantages for flexible grasping and manipulation than conventional grippers. Unlike a conventional jaw gripper with antipodal gripping force or a basic suction cup with vacuum gripping force, bio-inspired grippers often adapt to target objects via soft gripping pads and grasp objects with multiple grasp principles. For instance, Fig. 1 demonstrates the grasp principles of chameleon tongue and the related gripper that is inspired by the nature of the chameleon tongue and adopted in this work. It is composed of a soft gripping pad, which can deform to generate an airtight contact for a target object, and thus firmly grasp the object by multiple gripping forces, such as the vacuum gripper force and wrapping force.

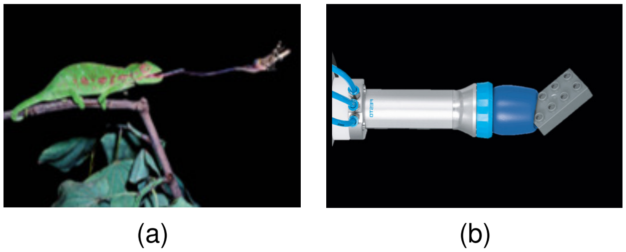

Fig. 1. Chameleon tongue and the related bio-inspired soft gripper [9]. (a) The flexible chameleon tongue that can adapt to the shape and size of a prey and firmly enclose the prey. (b) The investigated gripper inspired by the nature of the chameleon tongue.

Many studies report that deep neural networks trained on plenty of grasp examples can detect robust grasps for novel objects. However, few relevant works have been published on bio-inspired grasping based on deep learning, while a substantial amount of research exists on grasp planning with traditional jaw grippers and vacuum grippers using various neural networks [1]–[8].

To further boost the grasping performance and explore the potential for a bio-inspired gripper with deep learning, two main challenges have to be addressed: 1) A bio-inspired gripper could grasp objects with several grasp principles, hence a robust and effective neural network is required to learn the principles and estimate critical grasping parameters for the used gripper. 2) A massive dataset is needed to train the neural network. This letter addresses these challenges and proposes a framework for contact modeling, grasp scenario simulation, dataset collection and neural network development for a chameleon-tongue inspired soft gripper.

Fig. 2 shows the pipeline of the proposed grasping method, composed of two modules: the grasp evaluation and the grasp execution. The proposed neural network BIG-Net takes in a depth image of a grasp scene I and estimates the critical parameters for robotic grasping, including the grasp directions $\boldsymbol{D}_i$, the grasp qualities Q_i and the gripping steps S_i. A feasible grasp pose is computed based on the predicted grasp quality and direction. In addition, the real-time force-torque feedback of the gripper base is adopted to monitor the grasp status and optimize robot motion. More details of the grasp parameters are explained in Section III. The main contributions of this letter can be summarized as follows:

1) A grasp simulator is proposed for the bio-inspired soft gripper [9] to evaluate grasp robustness, including con-

Manuscript received: July 8, 2022; Revised November 7, 2022; Accepted: December 7, 2022. This paper was recommended for publication by Editor Hong Liu upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the China Scholarship Council under grant CSC201806090290. *(Corresponding author: Hui Zhang.)*

Hui Zhang and Karel Kellens are with the ACRO Research Group, Department of Mechanical Engineering, Wetenschapspark 27, 3590 Diepenbeek, Belgium, also with the Flanders Make ROB Core Lab@KU Leuven, 3000 Leuven, Belgium (e-mail: hui.zhang@kuleuven.be).

Yanming Wu and Eric Demeester are with the ACRO Research Group, Department of Mechanical Engineering, Wetenschapspark 27, 3590 Diepenbeek, Belgium.

Digital Object Identifier 10.1109/LRA.2022.3229237.

The source code, supplemental material, and video demo are available at https://github.com/huikul/BioInspiredGrasping.

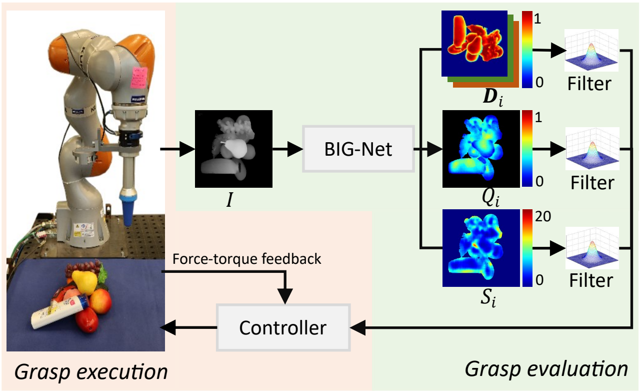

Fig. 2. Pipeline of the proposed grasping method. Note: 1) The images of $\boldsymbol{D}_i$ are visualized in the order of z-channel (at top layer with blue background), y-channel and x-channel. 2) The metric of S_i is mm. 3) Each grasp candidate is denoted as eight parameters, composed of a grasp pose in SE(3), a grasp quality and a gripping step. 4) The same visualization order for $\boldsymbol{D}_i$ and metric for S_i are used in Fig. 4, Fig. 7 and Fig. 8.

tact modeling, deformation tracking, gripping force estimation, etc. The grasp quality in simulation is estimated based on both the flatness of a contact surface and the mass center of a target object. A massive dataset is generated, containing **420 K** synthetic grasp scenes with more than **4.3 B** grasps in dense clutter.

2) A novel end-to-end neural network (BIG-Net) is proposed to estimate the pixel-wise grasp parameters for the applied bio-inspired gripper, including the grasp direction, grasp quality and gripping step.

3) A BIG-Net grasping method is presented and various experiments are implemented in both simulation and real world, including benchmarking tests with state-of-the-art methods for the random picking of novel objects in dense clutter. The proposed grasping framework works well for the chameleon-tongue inspired gripper.

The remainder of this letter is organized as follows: The related work is introduced in Section II. In Section III, concepts and construction of the proposed grasping method are described, followed by the elaborate procedures of the grasp simulation and network training in Section IV. Section V demonstrates extensive experiments with the proposed method and several baseline grasping methods. Finally, Section VI summarizes the performed work.

II. Related Work

A. Grasping with Bio-Inspired Soft Grippers

In the early stage, the research on grasping with bio-inspired grippers focused on developing multi-joint anthropomorphic hands and other grippers inspired by biological principles, such as the gecko-inspired adhesive [10], octopus-sucker effect [11] and chameleon tongue-sucker effect [9]. Novel bio-inspired grippers become more and more flexible with the development of shape-adaptive materials and under-actuated fingers [12].

Nevertheless, the grasping algorithms for bio-inspired grippers were improved slowly, despite various research on bio-inspired gripper designs being found. A lot of research investigated multi-sensor fusion and bio-inspired grasping, for instance, tactile sensing and object slip detection [13], but few focused on automatic grasping methods for bio-inspired grippers.

B. Deep Learning for Grasping

Automatic grasping methods have significantly evolved, resulting from the breakthrough technologies of both robotic perception and artificial intelligence in the last decades.

Abundant research reveals that neural networks trained on a large-scale dataset can learn to grasp. Neural networks with a 2D encoder-decoder architecture are able to detect feasible grasp poses effectively, benefiting from their abilities to take in a whole grasp scene and provide a global prediction of the grasp robustness. As an example, a GG-CNN was proposed by Morrison *et al.* [3], which runs a 50 Hz reactive grasping method. Similar frameworks include the GR-ConvNet [1] and SuctionNet-1Billion [7]. Furthermore, researchers reported that a neural network can learn to simplify the control scheme of a grasping task for a soft gripper [14]. Liu *et al.* [15] presented a deep reinforcement learning framework for multistage hybrid grasping with a novel soft gripper, achieving three grasping modes to deal with various objects.

C. Grasping Dataset

Huge grasping datasets are demanded to train neural networks for grasp planning, which can be generated by manual labeling [16]. Tedious work is needed to manually mark grasp examples. Alternatively, grasp examples can be collected by physical robots with various sensors to detect failed/successful grasps. However, a lot of robots and time are required in this approach [17].

In recent years, collecting synthetic grasping datasets in simulation has become popular [4], [5], [18]. The synthetic grasp examples are generated with virtual grippers and 3D meshes of objects in simulation, instead of physical robots and real-world objects. Numerous open-source synthetic datasets are available online to train a neural network for grasping, such as the Jacquard dataset [18] for parallel-jaw grippers and the Dex-Net 3.0 [6] for basic vacuum grippers. Furthermore, Lu *et al.* proposed a hybrid approach that collects grasp scenes by a physical camera and generates datasets in a simulator. These authors released two datasets for parallel-jaw grippers and vacuum grippers, named the GraspNet-1Billion [2] and SuctionNet-1Billion [7], respectively.

Nevertheless, many released datasets are merely compatible with traditional parallel-jaw grippers and vacuum grippers. Moreover, it is not easy to collect grasp examples for a bio-inspired soft gripper by manual labeling due to the flexible deformation of the soft gripping pad. Hence, collecting grasp examples by simulation is a proper method to generate a dataset for a bio-inspired gripper. Many existing simulation methods with simple human-designed grasp principles [7], [18] are not competent to create datasets for soft grippers, due to more complex contact surfaces than traditional grippers during grasping.

To address the problems above, a simulator is developed based on a previous framework of the authors [19], which

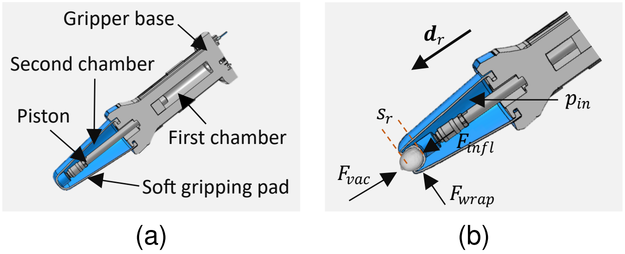

Fig. 3. (a) Main structure of the used gripper [9]. (b) Air pressure in the second chamber of the gripper p_{in}, grasp direction $\boldsymbol{d}_r$, gripping step s_i, and gripping forces F_{vac}, F_{wrap} and F_{infl}.

evaluates the grasp feasibility by tracking both the surface deformation and gripping force of the soft gripping pad in grasp scenes with stacked objects. The contact model in simulation is analyzed by a set of sub surfaces, instead of contact points in some published methods [6], [7].

III. Problem Formulation

In this work, the problem of robotic grasping is defined as predicting both the grasp feasibility and pose for novel objects from the depth image of a grasp scene and executing the grasp.

The adopted gripper $\mathcal{G}$ consists of a base frame and a double-acting cylinder filled with compressed air, as shown in Fig. 3 (a). The second chamber is fitted with a soft silicone gripping pad, which equates to the chameleon tongue. The piston is fastened to the silicone cap and closely separates the two chambers from each other. During the gripping procedure, the gripping pad touches an object and wraps the object with the moving of the piston, resulting in an airtight form fit and a strong holding force, as illustrated in Fig. 3 (b).

Instead of using a grasp representation based on the grasp center, tool rotation along the z-axis and required gripper width in [1], [3], this work denotes the grasp candidate in a robotic frame as an 8-parameter representation $\boldsymbol{g}_r$ in (1), where the grasp center $\boldsymbol{p}_r$ and the grasp direction $\boldsymbol{d}_r$ represent a grasp pose in SE(3), $q_r \subseteq (0,1]$ is the grasp quality to indicate the grasp robustness, $s_r > 0$ is the gripping step to ensure a tight contact between the gripping pad and the object surface, and their corresponding parameters in the image coordinate system (ICS) are denoted as $\boldsymbol{g}_i$, $\boldsymbol{p}_i$, $\boldsymbol{d}_i$, q_i, and s_i in (1). More details about these parameters are demonstrated in Fig. 3 (b) and further explained in Section IV-A. Similarly, all grasp candidates in the same grasp scene are described in (2), where T_{iw} converts pixels from the ICS to the world coordinate system (WCS), and T_{wr} is the transform matrix between the WCS and the robot coordinate system (RCS). Given a grasp scene $I \in \mathbb{R}^{1\times h_i \times w_i}$ with a size of $h_i \times w_i$, the pixel-wise grasp candidates can be defined by (3) with $\boldsymbol{G}_i \in \mathbb{R}^{8\times h_i \times w_i}$, $\boldsymbol{P}_i \in \mathbb{R}^{3\times h_i \times w_i}$, $\boldsymbol{D}_i \in \mathbb{R}^{3\times h_i \times w_i}$, $Q_i \in \mathbb{R}^{1\times h_i \times w_i}$ and $S_i \in \mathbb{R}^{1\times h_i \times w_i}$. In the matrices $\boldsymbol{G}_i$, $\boldsymbol{P}_i$, $\boldsymbol{D}_i$, Q_i and S_i, each element separately denotes the parameters $\boldsymbol{g}_i$, $\boldsymbol{p}_i$, $\boldsymbol{d}_i$, q_i, and s_i at the corresponding pixel in the grasp scene I.

Learning the end-to-end grasp robustness function BIG-Net in (4) is the main issue addressed in this paper, as the use of the force-torque feedback in Fig. 2 has been presented in our published work [8].

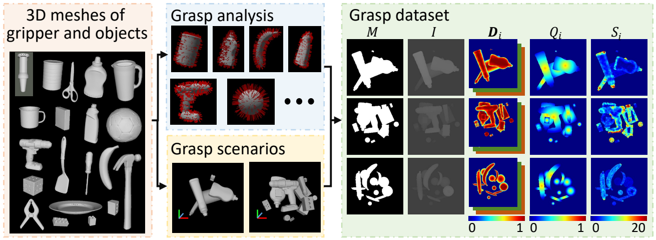

Fig. 4. Pipeline of the grasp simulation and dataset collection.

$$\boldsymbol{g}_r = (\boldsymbol{p}_r, \boldsymbol{d}_r, q_r, s_r), \qquad \boldsymbol{g}_i = (\boldsymbol{p}_i, \boldsymbol{d}_i, q_i, s_i) \tag{1}$$

$$\boldsymbol{G}_r = (\boldsymbol{P}_r, \boldsymbol{D}_r, Q_r, S_r) = T_{wr}(T_{iw}(\boldsymbol{G}_i)) \tag{2}$$

$$\boldsymbol{G}_i = (\boldsymbol{P}_i, \boldsymbol{D}_i, Q_i, S_i) \tag{3}$$

$$(\boldsymbol{D}_i, Q_i, S_i) = \text{BIG-Net}(I) \tag{4}$$

IV. Learning an End-to-End Grasp Robustness Function

The grasp robustness function BIG-Net aims to estimate $\boldsymbol{D}_i$, Q_i and S_i when a depth image I is given. The following subsections address two fundamental processes of the BIG-Net grasping method, including the collection of synthetic grasp scenes and the development of the BIG-Net.

A. Collection of Synthetic Grasp Scenes

The dataset for grasp estimation contains numerous depth images I with the corresponding "images" $\boldsymbol{D}_i$, Q_i, S_i and the binary mask images M that mark the foreground and background of I and are used to calculate the loss function during the network training.

Fig. 4 presents the pipeline of the grasp simulation to collect synthetic grasp scenes, working with three main steps: the contact modeling, the gripping force estimation and the grasp scene rendering. The following grasp simulation is implemented with two assumptions: 1) the target object $\mathcal{O}$ has an airtight surface and a rigid body with uniformly-distributed mass, 2) the gripping force is calculated based on Quasi-static physics with Coulomb friction.

1) Contact Modeling: Contact modeling and deformation tracking for $\mathcal{G}$ are the essential prerequisites to estimate the grasp feasibility in the simulation. The geometric dimension of $\mathcal{G}$ is defined as a frustum with the parameters $r_{\mathcal{G}}$, $R_{\mathcal{G}}$, H and l in Fig. 5 (a), where $r_{\mathcal{G}}$ and $R_{\mathcal{G}}$ respectively denote the head radius and the bottom radius of the gripping pad, H means the altitude of the frustum, and l represents the slant height of the frustum, named $l = \sqrt{(R_{\mathcal{G}} - r_{\mathcal{G}})^2 + H^2}$.

When the gripping pad contacts an object $\mathcal{O}$, part of the gripping pad will be located on the contact surface. The complex contact surface is simplified into a set of triangular sub surfaces $\mathbb{T} = \{t_1, \cdots, t_j, \cdots, t_n\}$, $\mathbb{T} \neq \emptyset$ and the corresponding vertices $\mathbb{V} = \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k, \cdots, \boldsymbol{v}_m\}$. Each vertex is defined as $\boldsymbol{v}_k(r, \theta, z)$ in the cylindrical coordinate system, considering the gripper base with a round frame. The model resolution is decided by the rotation step $\Delta\theta > 0$ and

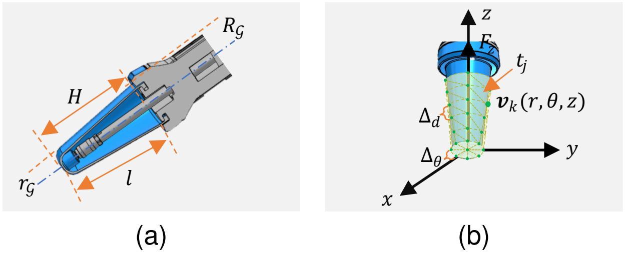

Fig. 5. Parameters of the used soft gripper. (a) Geometric dimension of the used gripper. (b) Coordinate system and parameters for the contact model.

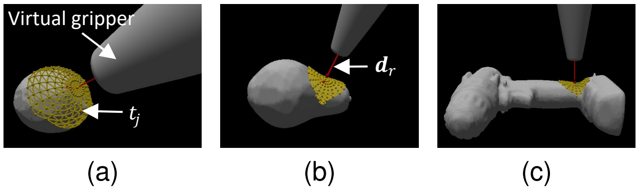

Fig. 6. Three virtual grasp trials and their contact models. Note: This figure simplifies the triangles for visualization, and more triangles are tracked for the dataset collection, depending on the parameters Δ_d, Δ_θ, etc.

the distance step $\Delta_d > 0$ in Fig. 5 (b). Neighbor points with the same rotation angle are defined by $\boldsymbol{v}_k(r_k, \theta, z_k)$, $\boldsymbol{v}_{k+1}(r_{k+1}, \theta, z_{k+1})$ and $r_{k+1} > r_k$ in the subsequent equations. When the gripper is moving along the -z direction in Fig. 5 (b) to wrap an object, the soft gripping pad is deformable under the constraints of its geometric and physical features, like the size, shape, and elasticity of the gripping pad. More than 10 rules are implemented to restrict the deformation of the gripping pad, and the main constraints are described in (5)-(8). Specifically, the equations (5)-(6) define the coordinate system and the dimension of the gripping pad in the simulation. The maximum distance between two neighbor points $\boldsymbol{v}_k(r_k, \theta, z_k)$ and $\boldsymbol{v}_{k+1}(r_{k+1}, \theta, z_{k+1})$ is constrained by (7), to avoid unrealistic deformation on the gripping pad. When the gripping pad touches the contact surface, the slope of the deformed gripping pads is never allowed to be larger than its slope at non-contact status, which is restricted by (8). Fig. 6 demonstrates three virtual grasps and their contact models.

$$\boldsymbol{n}_r = (1,0,0), \quad \boldsymbol{n}_z = (0,0,1), \quad 0 \leq \theta < 2\pi \tag{5}$$

$$0 \leq \boldsymbol{v}_k \cdot \boldsymbol{n}_r \leq R_{\mathcal{G}}, \; -H \leq \boldsymbol{v}_k \cdot \boldsymbol{n}_z \leq 0 \tag{6}$$

$$||\boldsymbol{v}_{k+1}(r_{k+1}, \theta, z_{k+1}) - \boldsymbol{v}_k(r_k, \theta, z_k)|| \leq \Delta_d \tag{7}$$

$$0 < \boldsymbol{v}_k(r_k, \theta, z_k) \cdot \boldsymbol{n}_r \leq r_{\mathcal{G}} + \frac{-z_k}{H}(R_{\mathcal{G}} - r_{\mathcal{G}}), \text{if } r_k \geq r_{\mathcal{G}} \tag{8}$$

2) Gripping Force Estimation: The grasp feasibility is related to the geometric and physical characteristics of a gripper. As for the applied gripper $\mathcal{G}$ in Fig. 3 (b) and Fig. 5 (b), it generates the gripping force F_z based on three main principles: 1) a vacuum gripping force F_{vac} exists, attributed to the airtight contact between the gripping pad and the object surface, 2) a wrapping force F_{wrap} holds the object due to the friction on the wrapping surface, and 3) an inflation force F_{infl} is generated towards the contact surface, resulting from the air pressure in the second chamber of $\mathcal{G}$.

Let $p_{air} > 0$ be the air pressure of atmosphere, $p_{in} > 0$ be the air pressure in the second chamber of the gripper, and $\mu > 0$ be the coefficient of friction on the contact surface. Briefly, the maximum gripping force $max(F_z)$ can be estimated by (9)-(12). Although it is difficult to conduct the peak values of F_{vac}, F_{wrap} and F_{infl} in the same grasping trial, $max(F_z)$ is a substantial reference to evaluate the grasp robustness.

$$max(F_{vac}) = p_{air}\pi R_{\mathcal{G}}^2 \tag{9}$$

$$max(F_{wrap}) = \mu p_{in}\pi l(r_{\mathcal{G}} + R_{\mathcal{G}}) \tag{10}$$

$$max(F_{infl}) = 0 \tag{11}$$

$$max(F_z) = max(F_{vac}) + max(F_{wrap}) + max(F_{infl}) \tag{12}$$

F_z can be analyzed by a set of sub forces on the sub gripping pads. Taking the sub surface $t_j \in \mathbb{T}$ as an example in Fig. 5 (b), $(n_x^{t_j}, n_y^{t_j}, n_z^{t_j})$ is the surface normal of t_j, and A^{t_j} is the area of t_j. Then, the vacuum gripping force, wrapping force and inflation force of t_j are respectively denoted as $f_{vac}^{t_j}$, $f_{wrap}^{t_j}$ and $f_{infl}^{t_j}$ in (13)-(15) with Quasi-static physics. As a result, F_z is the summary of all sub forces $f_{vac}^{t_j}$, $f_{wrap}^{t_j}$ and $f_{infl}^{t_j}$ in (16), which are related to flatness of a contact surface. Notably, the inflation forces $f_{infl}^{t_j}$ and F_{infl} always oppose the gripping force F_z based on the coordinate system in Fig. 5 (b). Hence, $max(F_{infl}) = 0$ and $f_{infl}^{t_j} < 0$ are respectively concluded in (11) and (15).

$$f_{vac}^{t_j} = p_{air}A^{t_j}n_z^{t_j} \tag{13}$$

$$f_{wrap}^{t_j} = \mu p_{in}A^{t_j}\sqrt{\left(n_x^{t_j}\right)^2 + \left(n_y^{t_j}\right)^2} \tag{14}$$

$$f_{infl}^{t_j} = -p_{in}A^{t_j}n_z^{t_j} \tag{15}$$

$$F_z = \sum(f_{vac}^{t_j} + f_{wrap}^{t_j} + f_{infl}^{t_j}), t_j \in \mathbb{T} \tag{16}$$

$$q = e^{-(1-\frac{F_z}{max(F_z)})}, q \subseteq (0,1] \tag{17}$$

The grasp quality metric is extremely important in the grasp simulation. Some existing work defines a threshold for the force-torque wrenches of grasping trials to evaluate good/bad grasps and predicts the grasp robustness via neural networks with classification architectures for real-world grasping. However, the threshold between "good grasp" and "bad grasp" is often unclear. It is hard to predict the grasp success when the force-torque wrench of a grasping trial nears the threshold. Therefore, the grasp feasibility is more appropriate to be calculated based on a quantitative metric.

In the proposed pipeline, a quantitative metric is defined for the grasp quality. It is related to the flatness of a contact surface and the distance between the grasp center and mass center of a target object. First, a basic grasp quality q for the contact model is denoted in (17), which is a normalized function of F_z. In this definition, a stronger F_z concludes a higher value of q. The nature of (17) is more sensitive when the value of q nears 1.0 but less sensitive for a low value of q. When this function is applied to collect a synthetic dataset, the BIG-Net trained on the dataset is more sensitive to detect high-quality grasp regions, which is helpful for grasping in the real world. Then, the final grasp quality metric q_i in grasp scenes is further explained by (20) in Section IV-A3.

3) Grasp Scene Rendering: Each grasp scene in the dataset consists of a depth image I with the corresponding “images” $\boldsymbol{D}_i$, Q_i, S_i and M. Over 50 3D meshes of objects from the YCB [20] and KIT [21] datasets are selected for the grasp simulation. Unlike many existing methods simulating each grasp scene with an isolated 3D mesh [5], [18], our simulator renders grasp scenes with stacked objects in clutter.

In detail, a virtual table $\mathcal{T}$ is defined on the plane $\prod(x, y, 0)$ in the WCS of a simulated grasp scene. Several 3D meshes are randomly selected, and their stable-pose sequences [4] are calculated. The 3D meshes are deployed on the virtual table with random stable poses from their pose sequences. Then, the subsequent batch of 3D meshes is randomly put upon the previous scene. To simulate a set of randomly stacked objects, the size of $\mathcal{T}$ is restricted, and a intersection-checking principle is followed for the interactions of the stacked objects.

A virtual camera $\mathcal{C}$ is deployed to render depth images I. In a depth image, each pixel $I(u, v)$ in the foreground is considered as the grasp center, named $\boldsymbol{p}_i(u, v)$. The corresponding grasp direction, grasp quality and gripping step are respectively denoted as $\boldsymbol{d}_i(u, v) \in \boldsymbol{D}_i$, $q_i(u,\ v) \in Q_i$ and $s_i(u, v) \in S_i$. Specifically, the grasp direction $\boldsymbol{d}_i(u, v)$ is computed based on the average normal of the contact surface, and the gripping step $s_i(u,\ v)$ depends on the boundary of the contact surface, as depicted in (18)-(19). The final grasp quality $q_i(u, v)$ is evaluated based on both the basic quality value q in (17) and the distance between $\boldsymbol{p}_i(u, v)$ and the mass center of the target object $\boldsymbol{c}_{\mathcal{O}}(u, v)$ in I. A closer distance contributes to a higher value of $q_i(u, v)$, which is more favorable to indicate the grasp robustness than many existing methods ignoring mass centers of objects. Let $\boldsymbol{p}_{\mathcal{O}}(u, v)$ be a random point on the target object $\mathcal{O}$ in the ICS, and $q_i(u, v)$ is formulated by (20).

$$\boldsymbol{d}_i(u, v) = \frac{1}{\sqrt{\sum (A^{t_j})^2}} \sum A^{t_j}(n_x^{t_j}, n_y^{t_j}, n_z^{t_j}),\ t_j \in \mathbb{T} \quad (18)$$

$$s_i(u, v) = -min(\boldsymbol{v}_k \cdot \boldsymbol{n}_z),\ \boldsymbol{v}_k \in \mathbb{V} \quad (19)$$

$$q_i(u, v) = q(u, v) \cdot (1 - \frac{||\boldsymbol{p}_i(u, v) - \boldsymbol{c}_{\mathcal{O}}(u, v)||}{max||\boldsymbol{p}_{\mathcal{O}}(u, v) - \boldsymbol{c}_{\mathcal{O}}(u, v)||}) \quad (20)$$

Based on the equations (5)-(20), plenty of grasp scenes with I, $\boldsymbol{D}_i$, Q_i, S_i and M are generated in the simulator. The 3D meshes of objects were rescaled with the factors of 0.5, 0.75 and 1.0 to simulate more objects with various sizes. More configurations of the simulator can be found in the supplemental material. Finally, 420 K depth images with randomly stacked objects were rendered and collected, containing over 4.3 B synthetic grasps (Fig. 4). Random noises with σ_I= 2 mm were added to the synthetic depth images to simulate the disturbances of a physical depth camera. 150 hours were consumed for the dataset collection on the PC mentioned in Section V-A. Table I compares basic information of the BIG-Net dataset and state-of-the-art grasping datasets, revealing that the proposed dataset is more similar to the real-world grasp scenes and contains more grasp examples.

TABLE I
BASIC INFORMATION OF THE PROPOSED DATASET AND FIVE STATE-OF-THE-ART DATASETS

Dataset	Modality	Method	Object's status	Image	Grasp
Dex-Net 2.0 [4]	Depth	Sim.	Isolated	6.7 M	6.7 M
GPD [5]	Depth	Sim.	Isolated	340 K	340 K
Jacquard [18]	RGB-D	Sim.	Isolated	54 K	1.1 M
SuctionNet -1Billion [7]	RGB-D	Sim., real	Multiple	97 K	1.1 B
BIG-Net	Depth	Sim.	Stacked	**420 K**	**4.3 B**

Note: Sim. is the abbreviation of Simulation.

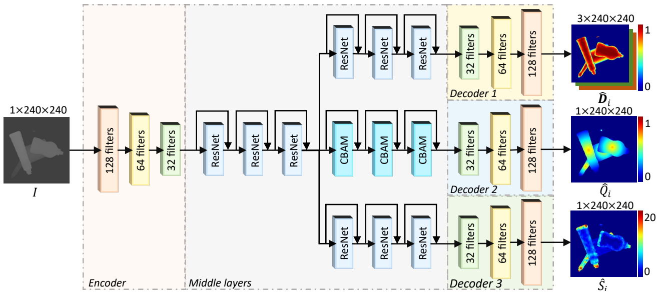

Fig. 7. Architecture of the BIG-Net. Note: The 3-layer encoder is composed of a 128-channel CNN with the 9×9 kernel, a 64-channel CNN with the 3×3 kernel and a 32-channel CNN with the 3×3 kernel. The decoders are defined with the same parameters by reverse order.

B. Development of Neural Network

1) Architecture: The function of the BIG-Net is to learn grasp principles and predict the grasp parameters during a physical grasping trial, named $(\boldsymbol{D}_i, Q_i, S_i)$ = BIG-Net(I). An encoder-decoder neural network is implemented for end-to-end grasp estimation, which takes in a depth image $I \in \mathbb{R}^{1\times240\times240}$ and provides three “images”, named $\boldsymbol{D}_i \in \mathbb{R}^{3\times240\times240}$, $Q_i \in \mathbb{R}^{1\times240\times240}$ and $S_i \in \mathbb{R}^{1\times240\times240}$.

Fig. 7 presents the architecture of BIG-Net. The BIG-Net extracts features from the input image via a 3-layer CNN encoder, followed by a series of 32-channel ResNets to encode the input image and learn the grasp principles in the middle layers. To decrease the BIG-Net parameters and train a condensed network, the middle layers involve three shared ResNet layers and three independent branches for the prediction of $\boldsymbol{D}_i$, Q_i and S_i. Notably, the branch of $\boldsymbol{D}_i$ in the middle layers is composed of revised CBAMs [22] to improve the robustness, while other branches consist of traditional ResNets. Finally, the three branches are separately decrypted by 3-layer CNN decoders for the output of $\boldsymbol{D}_i$, Q_i and S_i.

2) Training and Fine-Tuning: 400 K synthetic grasp scenes were used for the BIG-Net training, and 20 K grasp scenes were contained in the test dataset. Besides, the prediction error on the foreground of I, instead of the whole image, was taken into account for the backward propagation. The Mean Squared Error (MSE) loss function of the BIG-Net is defined as $L_{\text{BIG-Net}}$ in (21), where $\lambda_{\boldsymbol{D}}$, λ_Q and λ_S denote the weights of loss function at each branch. $L_{\boldsymbol{D}}$, L_Q and L_S are the MSE loss functions of $\boldsymbol{D}_i$, Q_i and S_i formulated in (22)-(24). After a few training trials, the learning rate was set as 0.0005, and a batch size of 80 was used. $\lambda_{\boldsymbol{D}} = \lambda_Q = \lambda_S = 1.0$ often reported better performance.

$$L_{\text{BIG-Net}} = \lambda_{\boldsymbol{D}} L_{\boldsymbol{D}} + \lambda_Q L_Q + \lambda_S L_S \tag{21}$$

$$L_{\boldsymbol{D}} = \text{MSE}(\hat{\boldsymbol{D}}_i \cdot M - \boldsymbol{D}_i \cdot M) \tag{22}$$

$$L_Q = \text{MSE}(\hat{Q}_i \cdot M - Q_i \cdot M) \tag{23}$$

$$L_S = \text{MSE}(\hat{S}_i \cdot M - S_i \cdot M) \tag{24}$$

$$\varepsilon_{\boldsymbol{d}} = arccos||\hat{d}_i \cdot d_i||, \ \varepsilon_q = |\hat{q}_i - q_i|, \ \varepsilon_s = |\hat{s}_i - s_i| \tag{25}$$

The network was fine-tuned by adjusting layers and channels of the sub modules in the BIG-Net. More than 40 similar networks were trained to find the optimal architecture for the grasp estimation, regarding the computational complexity t_I, and average prediction errors of the grasp direction, quality value and gripping step at the pixel wise in the foreground, named $\varepsilon_{\boldsymbol{d}}$, ε_q and ε_s in (25), wherein $\boldsymbol{d}_i$, q_i and s_i are the standard values in simulation (Section IV-A), and $\hat{\boldsymbol{d}}_i$, $\hat{q}_i$ and $\hat{s}_i$ are the predicted values via the BIG-Net.

Finally, the network model in Fig. 7 reported the best performance. It contains 2.41 M parameters and estimates the grasp robustness with $t_I = 6.34$ ms for a 240×240 depth image.

V. EXPERIMENTS

A. Setup

The proposed grasping method was assessed in both simulation and real environments with the setup shown in Section 1 of the supplemental material. The robotic system is composed of a 6-DoF cobot (KUKA LBR IIWA 14 R820), a bio-inspired soft gripper (FESTO DHEF-20-A), a wrist-mounted camera (Intel RealSense L515) and a PC. The PC running Ubuntu 20.04 OS was used in the experiments, which consists of a multi-kernel 3.5 GHz Intel Core i9-9920X CPU, 64 GB of dynamic system memory (DRAM), and two Nvidia GeForce RTX 2080Ti graphics cards.

B. Experiments in Simulation

This subsection demonstrates the experiments with synthetic grasp scenes. The $\boldsymbol{d}_i$, q_i, s_i, $\boldsymbol{D}_i$, Q_i and S_i in simulation are considered as the standard values and compared with the predicted values $\hat{\boldsymbol{d}}_i$, $\hat{q}_i$, $\hat{s}_i$, $\hat{\boldsymbol{D}}_i$, $\hat{Q}_i$ and $\hat{S}_i$ via the BIG-Net.

1) BIG-Nets Trained on Different Datasets: First, the performance of BIG-Nets was assessed at the pixel level. These tests aim to evaluate the robustness of BIG-Nets trained on different datasets and their performance for grasp scenarios with various sizes and objects' distributions. A new dataset containing 4.3 B grasps for **isolated** objects were rendered to train the BIG-Net I, using the pre-analyzed 3D meshes in Section IV-A. The BIG-Net II was trained on the grasp scenes with **stacked** objects mentioned in Section IV-A3. Both the BIG-Net I and BIG-Net II were trained on grasp scenes of 240×240 pixels. Table II lists two BIG-Nets and their average prediction errors in 10 K grasp scenes with stacked objects, named $\varepsilon_{\boldsymbol{d}}$, ε_q and ε_s denoted in (25). The 10 K grasp scenes with 102 M grasps were also synthesized with the pre-analyzed 3D meshes, which can be classified into three scenarios based on the size of a grasp scene and objects' distribution. Note that although the BIG-Nets were trained only with scenes of 240×240 pixels, images of Scenario 2 and Scenario 3 in the test dataset have a resolution of 480×480 pixels.

TABLE II
PREDICTION ERRORS OF TWO BIG-NETS

Grasp scenario		Scenario 1	Scenario 2	Scenario 3
Size of I (pixels)		240×240	480×480	480×480
Size of grasp scene (m^2)		0.35×0.35	0.70×0.70	0.70×0.70
Number of objects		20	40	80
Distribution (objects/m^2)		163	82	163
BIG-Net I	$\varepsilon_{\boldsymbol{d}}$	17.3°	12.9°	17.4°
	ε_q	0.192	0.166	0.200
	ε_s	0.6 mm	0.4 mm	0.4 mm
BIG-Net II	$\varepsilon_{\boldsymbol{d}}$	6.86°	6.29°	6.87°
	ε_q	0.085	0.083	0.083
	ε_s	0.3 mm	0.4 mm	0.3 mm

Note: The BIG-Net I and BIG-Net II were trained on synthetic grasp scenes with isolated objects and stacked objects, respectively.

Briefly, BIG-Nets often achieve similar prediction errors for grasp scenes with the same density of objects' distribution, e.g., Scenario 1 and Scenario 3, and they have better prediction accuracy when objects are less stacked in grasp scenes, e.g., Scenario 2. The BIG-Net I has lower accuracy in the aspects of $\hat{\boldsymbol{d}}_i$ and $\hat{q}_i$. On the contrary, the BIG-Net II can learn more about the grasp prediction from clutters and improve the robustness to deal with grasp scenes containing many covered objects. It reports the lower prediction errors of $\varepsilon_{\boldsymbol{d}} < 7.0°$, $\varepsilon_q < 0.09$ and $\varepsilon_s \leq 0.4$ mm, which outperforms the BIG-Net I. The $\varepsilon_{\boldsymbol{d}}$ of Scenario 2 is relatively lower, resulting from the sparser objects than in other scenarios. Additionally, the prediction of s_i is a fairly easy task, so ε_s does not significantly increase in Scenario 1 and Scenario 3. Consequently, the BIG-Net II has higher prediction accuracy in dense clutter, and it is used by default in the subsequent experiments.

2) Grasp Estimation for Synthetic Grasp Scenes: The error distribution of the BIG-Net II prediction was evaluated with synthetic depth images. Fig. 8 illustrates three synthetic grasp scenes with stacked objects from the test dataset (Section IV-B2) and their prediction errors $\varepsilon_{\boldsymbol{D}}$, ε_Q and ε_S that are calculated with the similar principles in (25).

As shown in Fig. 8, the BIG-Net II typically reports low prediction errors on $\varepsilon_{\boldsymbol{D}}$, ε_Q and ε_S when pixels are located on top-layer objects and far away from objects' boundaries, since the features of them are clearly presented in a depth image, and the BIG-Net II can effectively predict the grasp robustness. For the same reason, the BIG-Net II has poor performance on the boundaries between stacked objects and on lower-layer objects that are partly covered by others.

Fortunately, the inaccurate prediction of the BIG-Net II on the boundaries and covered objects does not cause a trouble for real-world grasping, because a feasible grasp pose with a high value of q_i in the real world is often located near to an object center. Furthermore, the grasping sequence for a clutter usually starts with a top-layer object by implementing a simple grasp

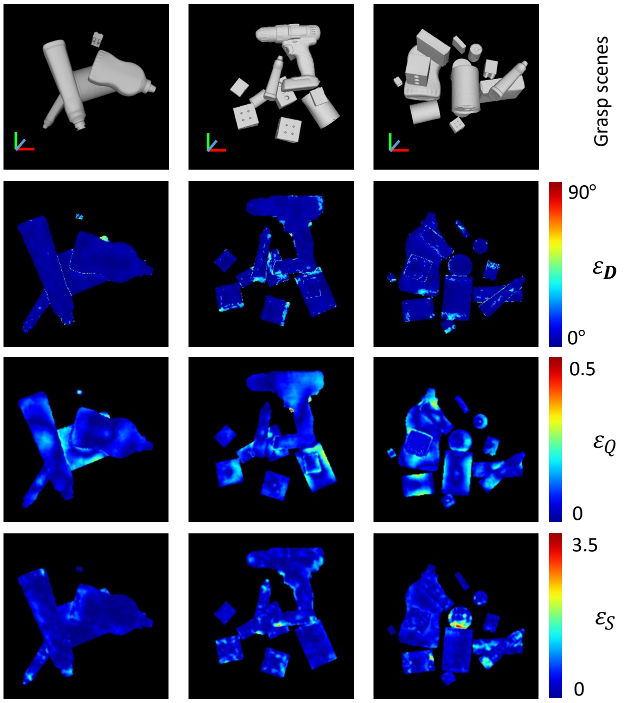

Fig. 8. Three synthetic grasp scenes and the corresponding ε_D, ε_Q and ε_S.

scene segmentation algorithm [8]. Once the top-layer objects are removed, the residual objects will become uncovered.

C. Experiments in Real World

This subsection demonstrates several benchmark experiments with the BIG-Net II and state-of-the-art grasping methods for real-world grasping. A dataset of 49 rigid, non-porous and novel objects were selected for physical grasping experiments. The objects are separated into two categories: 1) common household items (including 35 objects) [1], [3], [6], [7], and 2) adversarial items (including 14 objects from the Dex-Net dataset [4]). The random clutters for the experiments were deployed following the method proposed by ten Pas *et al.* [5], as presented in the supplemental material.

1) Picking with Different Grasping Principles: Fig. 9 illustrates four robotic grasp examples with different grasping principles for the related soft gripper. Fig. 9 (a) presents a flat object with an airtight surface, which is too large to be wrapped by the gripping pad. The BIG-Net II detects an optimized grasp pose at the center of the object surface, and the gripper picks up the object due to F_{vac} on the airtight contact surface. Furthermore, the BIG-Net II can find feasible grasp poses for objects without flat surfaces and conclude successful grasping trials by F_{wrap} and F_{infl}, e.g., Fig. 9 (b) and (d). With the varying of object poses, the BIG-Net II evaluates the object surface and decides the final grasp pose with different grasping principles. For instance, Fig. 9 (c) presents a coffee cup grasped by F_{vac}, and Fig. 9 (d) shows the same cup with another random pose, which is grasped by F_{wrap} and F_{infl}. All successful cases above indicate that the three gripping forces F_{vac}, F_{wrap} and F_{infl} in Section IV-A are essential and functional for the used soft gripper, and the BIG-Net II can learn the grasping principles effectively.

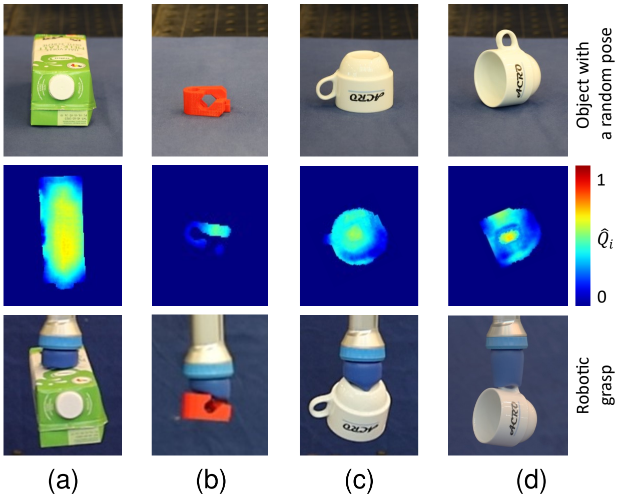

Fig. 9. Robotic grasping of four objects. (a) A flat object grasped by F_{vac}. (b) An adversarial object from the Dex-Net dataset grasped by F_{wrap} and F_{infl}. (c) A coffee cup grasped by F_{vac}. (d) The coffee cup in the subfigure (c) grasped by F_{wrap} and F_{infl}.

2) Random Picking of Novel Objects: The performance of BIG-Nets was in-depth investigated with traditional grasping methods. Four metrics were used to detect the grasp pose, as listed in Table III: 1) a random grasp point on a clutter, 2) a grasp point nearest to the clutter's center, 3) the BIG-Net I , and 4) the BIG-Net II. Table III shows that the traditional centroid grasping merely works well for grasp scenes with isolated objects. It reports a relatively lower success rate for adversarial items, because the feasible grasp region of an adversarial item is often not at its center, e.g., Fig. 9 (b). As a comparison, the BIG-Net grasping methods perform an average success rate of $> 80\%$ for objects in clutter, which typically consume 25 ms to detect a feasible grasp pose for a real-world grasp scene with a size of 40 cm×40 cm. Although the BIG-Net I performs the highest success rates for both isolated household and adversarial items, another one trained on stacked objects accomplishes the success rates of 94% for stacked household items and 86% for stacked adversarial items, respectively.

Furthermore, Table IV demonstrates a series of benchmarking tests with the BIG-Net II and state-of-the-art methods. In the revised Dex-Net 3.0 and SuctionNet-1Billion in this table, their neural network were not re-trained to avoid skewing the results. The main idea of revisions is to adjust their configurations to fit the dimension and parameters of the investigated bio-inspired gripper. For instance, the gripping step s_r is not provided by either the Dex-Net 3.0 or SuctionNet-1Billion framework, but it is necessary for the bio-inspired gripper. In the revised versions, s_r is estimated based on the depth values of the selected grasp region in the depth image I, and optimized according to the real-time force-torque feedback on the gripper base measured by the KUKA IIWA [8].

Table IV indicates that the BIG-Net II with the bio-inspired gripper conducts higher grasp success rates, since the bio-

TABLE III
SUCCESS RATES OF THE BIG-NET GRASPING METHODS AND TRADITIONAL METHODS

Approach	$S.R.$ (household, %)	$S.R.$ (adversarial, %)
	Isolated / Stacked	Isolated / Stacked
Random	65 / <50	<50 / <50
Centroid	98 / <50	87 / <50
BIG-Net I	**100** / 87	**97** / 83
BIG-Net II	**100** / **94**	96 / **86**

Note: $S.R.$ is the abbreviation of Success Rate.

TABLE IV
SUCCESS RATES OF THE BIG-NET II AND STATE-OF-THE-ART GRASPING METHODS FOR THE GRASPING OF STACKED OBJECTS

Approach	Gripper	$S.R.$ (%) Household	$S.R.$ (%) Adversarial
Dex-Net 3.0	Vacuum	82	81
SuctionNet-1Billion	Vacuum	81	-
Revised Dex-Net 3.0	Bio-inspired	85	67
Revised SuctionNet -1Billion	Bio-inspired	89	72
BIG-Net II	Bio-inspired	**94**	**86**

inspired gripper has a flexible gripping pad to fit an object, especially an adversarial object. The revised Dex-Net 3.0 and SuctionNet-1Billion can estimate the grasp quality Q_i for the bio-inspired gripper, because the vacuum gripping force F_{vac} is one of the grasp principles for the gripper. However, the success rates of the revised Dex-Net 3.0 and SuctionNet-1Billion decrease for adversarial items and are lower than the BIG-Net I in Table III, as the bio-inspired gripper grasps an object with two extra forces, named F_{wrap} and F_{infl}, which cannot be estimated via either the Dex-Net 3.0 or SuctionNet-1Billion. In contrast, the BIG-Net trained on stacked objects outperforms others in two grasp cases and concludes an average success rate of 90% in more than 300 grasping trials. Failure cases and limitations of the BIG-Net grasping method have to be reported in the supplemental document.

VI. CONCLUSION

In this letter, an automatic grasping method has been proposed for the chameleon-tongue inspired soft gripper, including grasp scenario simulation, dataset collection and neural network development. More than 420 K grasping scenes with 4.3 B grasps are collected in the simulation with stacked objects. The BIG-Net is developed based on an encoder-decoder architecture and CBAMs to learn the grasp principles. The proposed grasping method typically consumes 25 ms to detect a feasible grasp pose in the real world, and it achieves the success rates of 94% for the random picking of household items and 86% for that of adversarial items, which outperforms traditional and state-of-the-art methods.

Future work will include the improvement of the BIG-Net to estimate more parameters for robotic grasping. Moreover, the performed framework will be extended to fit more types of bio-inspired grippers.

REFERENCES

[1] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2020, pp. 9626–9633.

[2] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "GraspNet-1billion: A large-scale benchmark for general object grasping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11 444–11 453.

[3] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *Int. J. Robot. Res.*, vol. 39, no. 2-3, pp. 183–201, Jun. 2020.

[4] 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 *Robot.: Scien. Sys.*, Jul. 2017.

[5] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *Int. J. Robot. Res.*, vol. 36, no. 13-14, pp. 1455–1473, Oct. 2017.

[6] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, "Dex-Net 3.0: computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2018, pp. 5620–5627.

[7] H. Cao, H.-S. Fang, W. Liu, and C. Lu, "SuctionNet-1billion: A large-scale benchmark for suction grasping," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 8718–8725, Oct. 2021.

[8] H. Zhang, J. Peeters, E. Demeester, and K. Kellens, "A CNN-based grasp planning method for random picking of unknown objects with a vacuum gripper," *J. Intell. Robot. Syst.*, vol. 103, no. 64, pp. 1–19, Nov. 2021.

[9] Festo, "Adaptive shape gripper: EHEF-20-A," *Festo.com*, https://www.festo.com/us/en/a/8092533 [Accessed Apr. 28, 2022].

[10] H. Jiang, E. W. Hawkes, C. Fuller, M. A. Estrada, S. A. Suresh, N. Abcouwer, A. K. Han, S. Wang, C. J. Ploch, A. Parness, and M. R. Cutkosky, "A robotic device using gecko-inspired adhesives can grasp and manipulate large objects in microgravity," *Sci. Robot.*, vol. 2, no. 7, Jun. 2017.

[11] T. Tomokazu, S. Kikuchi, M. Suzuki, and S. Aoyagi, "Vacuum gripper imitated octopus sucker-effect of liquid membrane for absorption-," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Sep. 2015, pp. 2929–2936.

[12] T. Hassan, M. Manti, G. Passetti, N. d'Elia, M. Cianchetti, and C. Laschi, "Design and development of a bio-inspired, under-actuated soft gripper," in *Proc. IEEE Ann. Int. Conf. Eng. Med. Biol. Soc.*, Aug. 2015, pp. 3619–3622.

[13] A. Nakagawa-Silva, N. V. Thakor, J.-J. Cabibihan, and A. B. Soares, "A bio-inspired slip detection and reflex-like suppression method for robotic manipulators," *IEEE Sensors J.*, vol. 19, no. 24, pp. 12 443–12 453, Dec. 2019.

[14] D. S. Diaz Cortes, G. Hwang, and K.-U. Kyung, "Imitation learning based soft robotic grasping control without precise estimation of target posture," in *Proc. IEEE Conf. Soft Robot.*, Apr. 2021, pp. 149–154.

[15] F. Liu, B. Fang, F. Sun, X. Li, S. Sun, and H. Liu, "Hybrid robotic grasping with a soft multimodal gripper and a deep multistage learning scheme," 2022, *arXiv:2202.12796*.

[16] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, no. 4-5, pp. 705–724, Mar. 2015.

[17] 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," *Int. J. Robot. Res.*, vol. 37, no. 4-5, pp. 421–436, Jun. 2017.

[18] A. Depierre, E. Dellandrea, and L. Chen, "Jacquard: a large scale dataset for robotic grasp detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, Oct. 2018, pp. 3511–3516.

[19] H. Zhang, J. Peeters, E. Demeester, and K. Kellens, "Deep learning reactive robotic grasping with a versatile vacuum gripper," *IEEE Trans. Robot.*, 2022, DOI:10.1109/TRO.2022.3226148.

[20] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: using the Yale-CMU-Berkeley object and model set," *IEEE Robot. Autom. Mag.*, vol. 22, no. 3, pp. 36–52, Sep. 2015.

[21] 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," *Int. J. Robot. Res.*, vol. 31, no. 8, pp. 927–934, May 2012.

[22] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.