2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Gravity-aware Grasp Generation with Implicit Grasp Mode Selection for Underactuated Hands

Tianyi Ko*†, Takuya Ikeda*, Thomas Stewart*, Robert Lee*, Koichi Nishiwaki*

***Abstract*— Learning-based grasp detectors typically assume a precision grasp, where each finger only has one contact point, and estimate the grasp probability. In this work, we propose a data generation and learning pipeline that can leverage power grasping, which has more contact points with an enveloping configuration and is robust against both positioning error and force disturbance. To train a grasp detector to prioritize power grasping while still keeping precision grasping as the secondary choice, we propose to train the network against the magnitude of disturbance in the gravity direction a grasp can resist (gravity-rejection score) rather than the binary classification of success. We also provide an efficient data generation pipeline for a dataset with gravity-rejection score annotation. Evaluation in both simulation and real-robot clarifies the significant improvement in our approach, especially when the objects are heavy.**

I. Introduction

The majority of state-of-the-art machine learning-based grasp detectors [1]–[8] assume that each rigid finger only has one contact with a small contact region. Although such a grasp is easy to handle in both data generation and learning aspects, its limited contact region makes the grasp fragile. A typical failure mode with such grasp mode by a parallel-jaw gripper is that the object rotates around the axis connecting the two contact points, during which a translational displacement also occurs, resulting in the object dropping. The small contact region also imposes a high demand for accurate hand-eye calibration and hand positioning, which is challenging for low-cost manipulators or those under low-impedance control for safety.

Napier [9] pointed out that in order to firmly grasp an object, humans utilize the palm, and referred to such grasps as a *power grip*. Cutkosky [10] further extended the taxonomy of [9] in the robotics context and defined *power grasp* as grasps distinguished by large areas of contact between the grasped object and the surfaces of the fingers and palm, while classifying grasps where the object is held with the tips of the fingers and thumb as *precision grasp*. Despite the fact that there are multiple definitions of *power grasp*, e.g., Zhang et al. [11] redefined it as a type of grasp that its mechanism can resist passively against external forces without relying on feedback control of joint torques, and Zhang et al. [12] redefined it as a grasp with zero or less than zero connectivity, they share the same idea that utilizing multiple contacts of the intermediate finger links and palm improves grasp quality. In this paper, we introduce a data generation and learning framework to learn *power grasps* (with the definition of [10]) in addition to the common *precision grasps*, as shown in Figure 2 (a).

Fig. 1. While existing works only handle precision grasping, power grasping is more robust against both initial position error and post-grasp force disturbance. We propose a data generation pipeline and neural network model that prioritizes power grasping if applicable (left) but implicitly switches back to precision grasping if power grasping is not available due to collision with the table (middle) or other objects (right.)

Underactuated hands [13] have a larger number of joints than actuators. This passive compliance allows the hand to adapt to various objects with non-typical geometries without explicitly controlling each joint. Since it makes power grasping easier, this work focuses on utilizing underactuated hands, or more specifically, Robotiq 2F-85 [14]. However, such mechanically automatic property makes learning power grasping even challenging because whether the grasp is a power grasp or a precision grasp cannot be explicitly controlled but only decided through the interaction between the object and the hand's link mechanism. In addition, since a power grasp requires the fingers to wrap around the object, which is not always possible due to collisions with the environment, such as the tabletop or other objects (see Fig. 1 middle and right.) To this end, we propose a neural network architecture that prioritizes power grasping if applicable but implicitly switches back to precision grasping if power grasping is not available. More specifically, while prior works [1]–[8] only estimate the success probability of a grasp, we handle two values, namely (i) *gravity-rejection score* that represents the magnitude of disturbance in the gravity direction a grasp can support, which implicitly encourages power grasping since our data generation pipeline labels higher score for power grasps, and (ii) *grasp validness* which implicitly rejects grasps where the fingers will collide with the environment or it is too big for the hand to grasp.

Our contributions are as follows:

- We propose to measure the grasp quality by *gravity-*

∗ T. Ko, T. Ikeda, T. Stewart, R. Lee, K. Nishiwaki are with Woven by Toyota, Inc., 3-2-1 Nihonbashi-Muromachi, Chuo-ku, Tokyo, Japan.
† Corresponding author. tianyi.ko@woven.toyota

rejection score, which is the magnitude of disturbance in the gravity direction a grasp can support in the simulation.

- We provide a data generation / learning pipeline that can generate power grasps while automatically switching to precision grasping when power grasping is not available due to collision.
- We propose to evaluate a grasp detector with varying object weight in order to examine both the detection accuracy and robustness of the detected grasp.

II. Related Works

A. Training Data for Learned Grasp Detectors

Grasp quality metrics are key in generating synthetic training data for learning grasp detectors. An often selected approach is based on an antipodal grasp [15], adopted in [1], [2], [8]. By definition, though, it only supports antipodal grasps and is thus not applicable to power grasps with more than two contact points. Ferrari and Canny [16] proposed Q1 criteria, which is the maximum radius of a $\mathbb{R}^6$ sphere in the contact wrench space (GWS), with its center aligned with the origin, under the constraint that the L1 norm of the contact forces is one. Frequently referred to as ϵ-metrics, it is adopted in many works such as [6], [8], [17]–[20]. Weisz and Allen [21] discussed the robustness of the ϵ-metrics against the grasp pose error. As the ϵ-metrics' assumption that each contact can exert arbitrary reaction force within the friction cone limited its application, Winkelbauer et al. [19] extended ϵ-metrics to consider the torque constraint of articulated fingers. Nevertheless, the constraint of the ϵ-metrics on the reaction force L1 norm prevents it from being adopted for power grasps that utilize kinematic constraint forces. To provide a quality metric for power grasps, Mirzal et al. [22] discussed the stability region of a power grasp. Zhang et al. [11] proposed a virtual work-based quality measure. However, although these works assume fully actuated hands, many robot hands that support power grasping (including ours) have an underactuated mechanism, making it hard to apply those metrics directly.

While analytical metrics are computationally efficient, it is difficult to capture all aspects of a contact, such as contact force distribution or motion after a slip. Kappler et al. [23] reported that a simulation-based approach achieves a better performance. Zhou et al. [24] and Eppner et al. [25] followed [23] by shaking the hand in a gravity-less simulation. Breyer et al. [4] and Jiang et al. [5] evaluated grasps by actually lifting the object in a simulation and classified its success or failure. Huang et al. [7] further shake the hand after lifting the object to select only robust grasps. Those prior works use the simulator to get a set of valid grasps. However, such binary classification suffers from the sim-to-real gap since the absolute value of success or failure highly depends on the simulator's contact model and parameters. In this paper, we also take a simulation-based approach, but we acquire a continuous spectrum of the *gravity-rejection score* to train the neural network to prioritize grasps with higher safety margins. A similar idea can be found in Fang et al. [20],

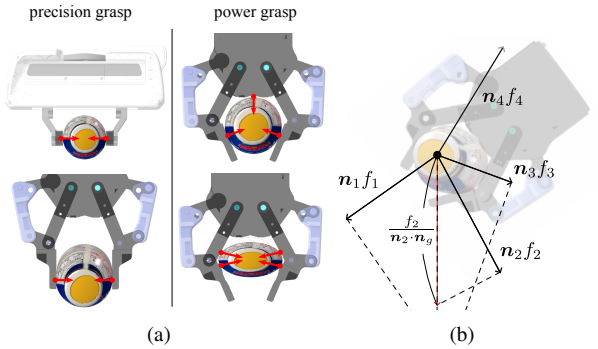

Fig. 2. (a) Most works assume a precision grasp (left) due to its simplicity. A power grasp (right) allows a more robust grasp thanks to the larger number of contacts. (b) Illustration of Eq. 1 which approximates the multidimensional *disturbance-rejection score* to the gravity direction *gravity-rejection score*.

where a score based on the distance from the object's center of mass was used, but our approach considers more diverse cases such as power grasping. In addition, we propose an approximation to allow the two-staged approach proposed by Qin et al. [1] to relax the high computation cost of the simulation-based approaches. Section III details our data generation pipeline.

B. Grasp Representation in Learned Graps Detectors

Grasp representation is another key to learning grasp detectors. For 2D top-down grasps, the grasps were typically represented as the hand's x-y location and yaw rotation on the height map [17], [26]. This approach is difficult to apply to power grasping because the hand's insertion depth cannot be heuristically derived. For grasp detection in 3D space, Ten Pas et al. [27] first sampled points on the object surface and then searched for grasp candidates in the Darboux frame, followed by a learned grasp classifier. However, their "push" operation makes it difficult to detect both power grasps and precision grasps simultaneously. Qin et al. [1] regressed translation offset and rotation of the hand from each point on the object surface. Cai et al. [8] used the location and normal of the points on the object surface to decide five degrees of freedom of the hand and regressed the remaining rotation in the plane. Since [1], [8] interpreted the surface point as candidates facing the hand, a set of precision grasp and power grasp may be assigned to a single point, leading to representation ambiguity. Zhao et al. [2] detected the middle point of two contact points to reconstruct a grasp, thus only supporting antipodal grasps. Sundermeyer et al. [3] and Huang et al. [7] interpreted surface points as candidates of contact and reported superior performance. However, contact points of a power grasp significantly differ depending on object size, making explicit contact representation challenging (see Fig. 2 (a) where the contact points differ among different grasp modes.)

Wang et al. [6] first detected the "view" and evaluated grasps represented in the "view" coordinate with discretized hand depth and in-plane rotation. This approach is applicable to power grasping because a precision grasp and a power

grasp are assigned to different discretized hand depths (see Fig. 2 (a) where the distance from the object differs for the different grasp modes). Breyer et al. [4] and Jiang et al. [5] voxelized the workspace and detected per-voxel grasp probability and hand rotation. This representation is also applicable to detecting power grasps because (i) a precision grasp and a power grasp are assigned to different voxels, (ii) a precision grasp and a power grasp that grasps a similar location on the object tend to have a similar rotation so that we can expect a continuous grasp rotation field. In this work, we follow this voxel approach due to the simple architecture and match for power grasping, but train the network with the *gravity-rejection score*.

III. Synthetic Data Generation with Gravity-Rejection Score

Given a set of scenes (in this work, 5K) with multiple objects (in this work, 1-5) placed in a cluttered arrangement (see Fig. 3 bottom), our target is to generate a dense annotation (in this work, 1.2K on average for each scene) of grasps with both power grasp and precision grasp mode, and label the *gravity-rejection score*, which stands for the magnitude of disturbance in the gravity direction that the grasp can support.

Following the previous works [1], [2], [8], we first generate per-object grasp poses. We first generate precision grasps by sampling antipodal points on the object surface. In order to generate power grasps, we perturb precision grasps in translation and rotation and simulate the contact process by closing the hand against the object in a gravityless simulation. Through the contact, the object settles in the hand, and we then record the hand's pose relative to the object. This simulation-based approach is beneficial because we can avoid explicitly modeling the complex mechanism of underactuated fingers, where the final hand configuration is only decided through the interaction between the object and the spring-loaded closed-link mechanism. For each scene, we project the per-object grasp poses to the scene based on the object's pose, and remove those that collide with the table or other objects.

In order to train a network to prioritize power grasping, we need a grasp metric that takes on a higher value for power grasps. As discussed in Sec. II-A, we take a simulation-based approach rather than an analytical one [11], [15], [16], [22]. However, unlike the case of [4], [5], [7] where the grasps are classified as success or failure, we measure the maximum disturance in the gravity direction that a grasp can support in the simulation. We call this a *gravity-rejection score*, which has a continuous value with the unit of [N].

One drawback of the simulation-based approach is the computation time. In our case, the naive approach where each grasp in each scene is evaluated by a simulation will result in roughly 6M simulations with varying gravity. While this is not intractable, thanks to modern cloud parallel computing, this paper proposes an approximation to reduce the number of simulations. Specifically, we perform the simulation in the per-object stage rather than the per-scene stage. For each

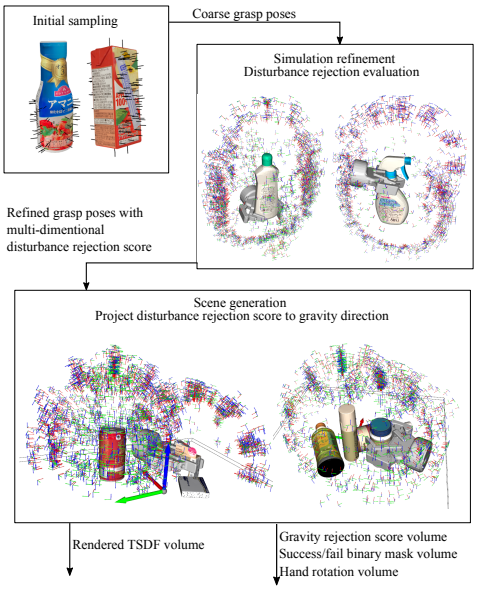

Fig. 3. Schematic of our data generation pipeline. We first sample antipodal grasps and randomize them (top-left). Those initial grasps are refined through grasp simulation (top-right). We apply external force in multiple directions and acquire a multi-dimensional *disturbance-rejection score* for each grasp. Finally, we create multi-object scenes and project per-object grasp poses to the scene to acquire the training data (bottom). The *disturbance-rejection score* is projected to the scene's gravity direction to acquire the *gravity-rejection score*.

grasp, in a gravity-free simulation, we apply an increasing external force to the object until it leaves the hand and then record the magnitude of that force. By applying this force to the object's center of mass, we can encourage the network to prioritize grasping near the center of mass, in addition to prioritizing power grasping.

Since the object's pose relative to the scene's gravity direction is unknown, we perform this evaluation in multiple directions, resulting in an n-dimentional *disturbance-rejection score* where n is the number of directions (in this work, 6). In order to project the n-dimentional *disturbance-rejection score* to the scene's gravity direction to acquire the one-dimensional *gravity-rejection score* f_g, we take the following approximation:

$$f_g = \min_{\{i|\boldsymbol{n}_i \cdot \boldsymbol{n}_g > \epsilon\}} \frac{f_i}{\boldsymbol{n}_i \cdot \boldsymbol{n}_g} \tag{1}$$

where $\boldsymbol{n}_g \in \mathbb{R}^3$ is the unit vector heading the direction of gravity in a scene, f_i is the i-th element of the *disturbance-rejection score*, $\boldsymbol{n}_i \in \mathbb{R}^3$ is the unit vector heading the direction of the i-th disturbance projected to the scene frame, and ϵ is a small positive value. Figure 2 (b) illustrates the approximation and Fig. 3 illustrates the whole data generation pipeline.

Thanks to the reduction in the number of simulations, we can employ accurate yet time-consuming simulation setups. Specifically, we use the hydroelastic contact model [28] implemented in Drake [29], which is highly realistic even with a concave-shaped object mesh. Simulation for all grasps for a single object typically takes roughly 12-18 hours

with a single AWS m5.2xlarge instance, but it is easy to parallelize since each simulation is independent. In addition, the computation time highly depends on the settings. For example, a coarse mesh speeds up the simulation in exchange for a less accurate object shape, a larger simulation step linearly improves the simulation time in exchange for a higher chance of failure, and simply switching the contact model to the point-contact results in a more than x10 faster but much more unstable simulation.

IV. Network Architecture and Training

Given a single depth image as the input, our target is to infer a set of grasp poses in SE(3) with *gravity-rejection score*. At the inference time, the robot tries to execute the first reachable and collision-free grasp with the highest *gravity-rejection score*. This differs from other works that estimate the probability of the grasp: a higher *probability* means the network is more *confident*, but it doesn't mean the grasp is more *robust*; in our case the *gravity-rejection score* has a unit of [N], which directly represents the grasp's robustness in the gravity direction.

As discussed in Section II-B, we take a 3D fully convolutional network as the backbone following [4], but with a *gravity-rejection score* regression head. Unlike the case of [4], where both positive and negative samples were included in the dataset, our data generation pipeline described in Sec. III only provides positive samples. We therefore train the *gravity-rejection score* head with L2 loss against positive examples only, leaving the invalid grasp regions as out-of-domain. While [4] introduced a heuristic filter to remove out-of-domain voxels, in this work, we propose to learn a *grasp validness* head to classify whether the voxel is out-of-domain. The evaluation in Sec. V-A shows that this classification head is also effective in collision avoidance or rejecting grasping an improper region. As the training data is densely annotated, we fill all voxels without a grasp label as *invalid* and use a weighted cross-entropy loss for the training.

V. Experiment and Evaluation

A. Analysis of Gravity-Rejection Score and Classification

This subsection provides a qualitative analysis of the network output *gravity-rejection score* and *grasp validness*. We placed two primitive objects in the simulator: a $\varnothing 65 \times 200$ mm cylinder at the location $[-0.0325, 0, 0]$ m and a $100 \times 80 \times 100$ mm cube at $[0.06, 0, 0]$ m. Even though these two objects were not included in our dataset, their geometries fall into our typical cases, such as water bottles and daily goods boxes, because we target home-use products. We placed a single camera at $[0.5, 0, 0.5]$ [m] and plotted the input TSDF volume and the output grasp volume. Figure 4 left illustrates the TCP (tool center point) coordinate, and Fig. 4 right shows the experiment scene. We plot the cross-sections by $z = 0.15$, $z = 0.08$ and $y = 0$ plane in Fig. 5, 6, 7 respectively.

The color bar on the right plot of the figures stands for the *gravity-rejection score*, which is a continuous field that takes on a higher value near the object. This distribution

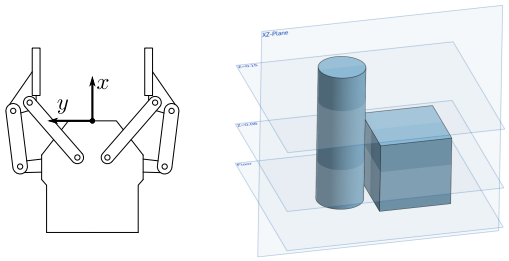

Fig. 4. TCP coordinate definition (left) and scene for the static network input/output analysis (right.) A $\varnothing 65 \times 200$ mm cylinder and a $100 \times 80 \times 100$ mm cube are placed on the surface and the input/output of the network is visualized with the three blue-colored cross-sections.

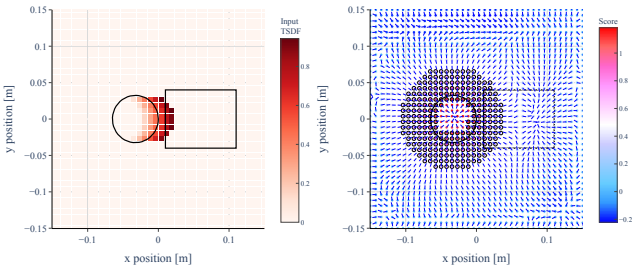

Fig. 5. Input TSDF volume (left) and output grasp volume (right) with $z = 0.15$ cross-section. The markers with black-circle are voxels with positive classification. The short blue lines are the hand approach direction.

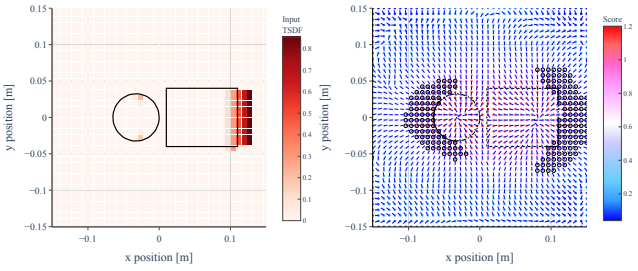

Fig. 6. Input TSDF volume (left) and output grasp volume (right) with $z = 0.08$ cross-section. The score prioritizes power grasps while the classification rejects grasps with collision or with too-big width.

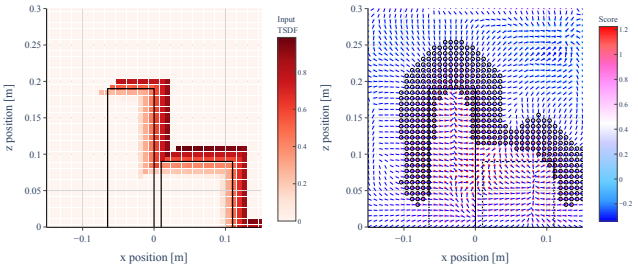

Fig. 7. Input TSDF volume (left) and output grasp volume (right) with $y = 0$ cross-section. Even for the region invisible to the camera, the network well predicts grasp poses. The classification properly removes grasps where the palm may collide with other objects.

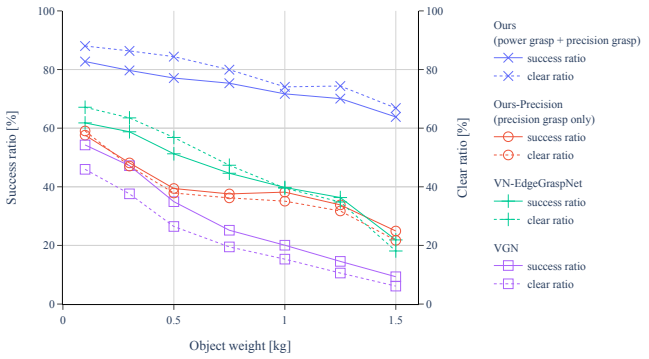

Fig. 8. Success ratio (SR) and clear ratio (CR) against multiple object weights. Our approach significantly outperformed the baselines, especially when the objects were heavy, showing the advantage of power grasp.

encourages power grasping, as shown in Fig. 2 (a). Since the *gravity-rejection score* head is only trained against valid grasps, its output for the out-of-domain voxels is incorrect, e.g., it takes on an even higher value inside the object and a medium score at regions far removed from the object. The markers surrounded by a black circle represent the voxel with a higher *grasp validness* than the threshold. The three figures show that this classification is highly effective in rejecting those too-close or too-far grasps. In addition, in Fig. 6, grasps at $x = 0$, $y = \pm 0.05$ region are rejected. Indeed, grasping this region is impossible due to the limitation of the hand's maximum aperture of 85 mm; thus, training data does not contain such grasps. Similarly, in Fig. 7 grasps at $x = 0.03$, $y = 0.12$ are rejected. This is also reasonable since the hand's base will collide with the objects in order to grasp this concave part.

B. Simulation Benchmark

In this subsection, we compare our approach (*ours*) with two baselines: Breyer et al. [4] (*VGN*) since we share the same backbone, and Huang et al. [7] (*VN-EdgeGraspNet*) since they reported state-of-the-art performance. While both [4] and [7] provide a pre-trained weight for Franka Emika's hand [30] [1] , we target Robotiq 2F-85 [14] hand; thus, we created a simulation model for both hands with an identical grasp force and contact parameters[2]. As the simulator, we used Drake [29] since its hydroelastic contact model [28] allows for a stable simulation of power grasping by hands with a passive parallel-link mechanism. To clarify the effect of power grasps, we also trained our model (*Ours-Precision*) for Franka Emika's hand with only precision grasp annotation in the training data.

For the quantitative evaluation, we used the same metrics as [1], [2], [4], [5], [7], [8]: success ratio (SR) representing the number of successful grasps divided by the number of total trials, and clear ratio (CR) representing the number of successful grasps divided by the total number of objects. Each scene is terminated after two consecutive grasp failures or detection failures. We created 128 scenes with the mesh models provided by Breyer et al. [4] [3] which are used by [5], [7] also.

While SR and CR are intuitive metrics, they mix two distinct aspects of the grasp quality: positional accuracy and robustness against disturbance. We, therefore, propose to evaluate grasp detectors against varying object weights since, in the lightweight region, they mainly capture positional accuracy, while in the heavy region, they capture both aspects. The ability to support a heavier object also means a grasp can lift the same objects with less grasp force, which is critical for handling fragile objects. Note that changing other physical properties, such as the friction coefficient, is another option, but it is harder to disentangle the two aspects. For example, lower friction requires both high positional accuracy (because only highly antipodal grasps are acceptable in precision grasping) and high robustness (because the object will slip more easily).

Figure 8 plots the SR and CR against the object weight from 0.1 kg to 1.5 kg. In the light-weight region (less than 0.5 kg), *Ours-Precision* performed similarly with *VGN* while *VN-EdgeGraspNet* outperformed both. This is reasonable because we share the same backbone as *VGN*, and the author of *VN-EdgeGraspNet* reported that their grasp representation is more advantageous. In the heavy-weight region (more than 1 kg), however, *Ours-Precision* outperformed *VGN* and performed on par with *VN-EdgeGraspNet*. This indicates that even when limited to precision grasping, our *gravity-rejection score* is beneficial in grasping heavy objects. One hypothesis is that if the hand only supports precision grasping, introducing the *gravigy-rejection score* to the edge-grasp representation will provide a better score. Nevertheless, our focus is on leveraging power grasping. *Ours*, with all features including power grasp annotation, significantly outperformed the others, with 20.9 % improvement for both SR and CR in the case of 0.1 kg and 41.9 / 48.7 % improvement of SR / CR in the case of 1.5 kg compared with *VN-EdgeGraspNet*.

It was surprising that even in the light-weight region, *Ours* outperformed the other cases. One reason is simply because the Robotiq 2F-85 hand has a slightly larger fingertip than that of the Franka Emika's hand, but we also observed that power grasping is more robust against position error. In a power grasp, the palm or the proximal finger link hits the object first, then the hand closes, and the distal fingers wrap around it. Even if the hand is inserted too deeply, the grasp is still valid as long as the object doesn't fall down in the case of a side grasp. Conversely, even when the hand insertion is too shallow, the distal fingers' wrap-around motion still "pulls" the object into the hand, or, in the worst case, results in a precision grasp.

Figure 8 also shows that *Ours* had the minimum drop of SR / CR due to the increasing object weight. This quantitatively proves the benefit of power grasping. Qualitatively,

[1] [7] performed a real-robot validation with the Robotiq gripper in their paper, but they only made Franka Emika version's weight public.

[2]We set 40 N for grasp force, 0.75/0.5 for static/dynamic friction coefficient, 5 for Hunt Crossley dissipation, 10^8 for hydroelastic-modulus.

[3]We merged the two test mesh sets for *packed* and *pile* and removed those that didn't fit our hand.

TABLE I

SUCCESS RATIO (SR) AND CLEAR RATIO (CR) UNDER DIFFERENT ROTATION REPRESENTATION

	Rotation Matrix ($\mathbf{e}_x$-$\mathbf{e}_z$)	Rotation Matrix ($\mathbf{e}_y$-$\mathbf{e}_z$)	Quaternion
SR [%]	**71.7**	68.0	57.8
CR [%]	**74.1**	69.2	61.0

we observed that power grasping frequently occurred for cylinder-shaped objects and that it was especially effective in lifting heavy cylinders laid down on the surface by a top-down grasp. We also observed that when the hand attempts to perform a power grasp on a thin object laid down on the surface, the grasp typically fails due to the fingertip's collision with the surface and the finger's undesired displacement due to passive compliance. However, such failure mode was rare, and in most cases, a precision grasp was selected. This shows that utilizing the *grasp validness* is effective in allowing the network to automatically fall back to precision grasping when power grasping is not available.

C. Comparison on the Rotation Representation

Following Qin et al. [1], we also select the rotation matrix as the rotation representation due to its continuity in the SO(3) [31]. However, [1] lacks a discussion on how to select the basis vectors of the rotation matrix, i.e., how we should define the TCP coordinate. A rotation matrix $R = [\boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z] \in \mathbb{R}^{3\times3}$ is constructed by three basis vectors. The network estimates two of them $\tilde{\boldsymbol{e}}_x, \tilde{\boldsymbol{e}}_z$ and reconstructs R by:

$$\begin{aligned} \mathbf{e}_x &= \tilde{\mathbf{e}}_x/|\tilde{\mathbf{e}}_x| \\ \mathbf{e}_z &= \mathbf{e}'_z/|\mathbf{e}'_z|,\ \mathbf{e}'_z = \tilde{\mathbf{e}}_z - (\mathbf{e}_x \cdot \tilde{\mathbf{e}}_z)\mathbf{e}_x \\ \mathbf{e}_y &= \mathbf{e}_z \times \mathbf{e}_x \end{aligned} \tag{2}$$

which means it prioritizes $\tilde{\boldsymbol{e}}_x$ to decide two dimensions of the rotation and supplementarily uses $\tilde{\boldsymbol{e}}_z$ to decide the remaining one, while doesn't explicitly estimate the $\boldsymbol{e}_y$ direction. We name this order as $\boldsymbol{e}_x$-$\boldsymbol{e}_z$ and compare it with the opposite order $\boldsymbol{e}_y$-$\boldsymbol{e}_z$, in addition to a quaternion version (see Fig. 4 left for the TCP coordinate definition.) Table I summarizes the result. In addition to the large gap between the rotation matrix and quaternion, we also observe an improvement of $\boldsymbol{e}_x$-$\boldsymbol{e}_z$ over $\boldsymbol{e}_y$-$\boldsymbol{e}_z$. This indicates that $\boldsymbol{e}_x$ is easier for the network to learn than $\boldsymbol{e}_y$. In the other part of this paper, we learn $\boldsymbol{e}_x$ and $\boldsymbol{e}_z$, and use $\boldsymbol{e}_x$-$\boldsymbol{e}_z$ order for the inference.

D. Validation with a Physical System

In this subsection, we validate our approach with a physical system, as shown in Fig. 9. Since we only have physical access to the Robotiq hand, we trained a baseline version of our model to mimic [4] by (i) training data containing only precision grasps, (ii) we only used *grasp validness* and ignored *gravity-rejection score* for the inference, (iii) quaternion rotation representation. We selected nine kinds of objects from the YCB dataset [32], which we didn't use for training, and created 20 scenes with 51 objects

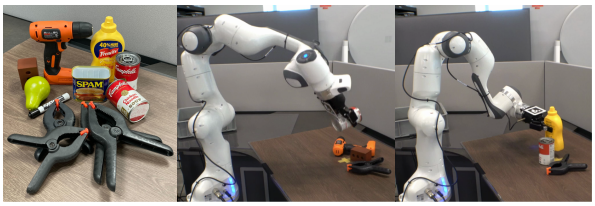

Fig. 9. Target objects (left) and capture (middle, right) of the validation.

TABLE II

SUCCESS RATIO AND CLEAR RATIO IN THE REAL ROBOT VALIDATION

	Objects	Trials	Success	SR [%]	CR [%]
Ours	51	51	38	**74.5**	**74.5**
Baseline	51	48	31	64.6	60.8

in total in a simulator. In [4], the objects were put in a box and dumped on the table. However, since our target objects were often rigid and round-shaped, they frequently rolled out of the workspace. Thus, we first created the scenes with the simulator, overlaid the scene's mesh with the real-time-streamed point cloud, and manually adjusted the objects' poses in the real world. Depth image was acquired from a pair of RGB cameras and a learning-based stereo inference [33]. Table II summarizes the result.

The absolute number of SR / CR was lower than the value reported by [4]. We assume this is because of the different object properties. In the case of [4], most objects were light and soft toys, which are relatively easy to grasp. Indeed, our objects also contained such an object (`061_foam_brick`), and we observed that in any case, the grasp succeeds. However, our objects contained more heavy and rigid objects. For example, the unopened `005_tomato_soup_can` weighed 350 g, and its hard metal exterior made the contact region small. The object even harder to grasp was `035_power_drill`, with more than 600 g weight and unevenly distributed mass. Other factors that may have contributed to the difference include the camera system, hand-eye calibration, and the underlying robot controller. Similarly, due to the sim-to-real gap, we cannot compare the absolute SR / CR with the simulation evaluation in V-B.

In order to avoid such biases, our evaluation was performed under the exact same system configuration, and our approach significantly outperformed the baseline. This result is consistent with the simulation and further justifies the effect of our contribution.

VI. LIMITATION

In this work, we reconstruct the hand pose through the location of the voxel corresponding to the TCP and the regressed rotation. This representation, though, is fragile against the rotation regression error, since a small error in the rotation results in a large error in the contact regions. We frequently observed failure cases in which the hand approach direction was inaccurate, causing the hand to grasp at an improper position (sometimes even not making contact with

the object at all). Another major failure case was when the finger collided with the object during the approaching motion due to an incorrect hand yaw rotation. Our attempt in Sec. V-C is one countermeasure for such a problem, but there is still a large margin for improvement. Contact-based grasp representations [3], [7] are more robust to rotation error. Finally, it is difficult to generate power grasps due to the inconsistent contact region among different grasp modes, even though our *gravity-rejection score* is also applicable to them. An approach to detect power grasps through contact-based representations remains in our future work.

VII. Conclusion

In this paper, we presented a data generation and learning framework to leverage power grasping in addition to precision grasping. We proposed to train a neural network to predict a continuous *gravity-rejection score* which is the magnitude of disturbance in the gravity direction that a grasp can support. Unlike the grasp probability, this allows the network prioritize power grasping while still keeping precision grasping as a secondary choice. We also proposed a data generation pipeline to efficiently create a dataset with *gravity-rejection score* annotation. We provided a qualitative analysis of the network output and showed the effect of *gravity-rejection score* in generating power grasps. For the quantitative evaluation, we proposed to evaluate the performance against varying object weights. A simulation evaluation proved a significant improvement in our approach, where power grasping improves the grasp robustness to positioning accuracy error and gravity direction force disturbance. Finally, we provided real robot validation, which showed that our approach is also effective in a physical system.

References

[1] 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 *Conf. on Robot Learning*. PMLR, 2020, pp. 53–65.

[2] B. Zhao, H. Zhang, X. Lan, H. Wang, Z. Tian, and N. Zheng, "Regnet: Region-based grasp network for end-to-end grasp detection in point clouds," in *Int'l. Conf. on Robotics and Automation*. IEEE, 2021, pp. 13 474–13 480.

[3] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *Int'l. Conf. on Robotics and Automation*. IEEE, 2021, pp. 13 438–13 444.

[4] M. Breyer, J. J. Chung, L. Ott, R. Siegwart, and J. Nieto, "Volumetric grasping network: Real-time 6 dof grasp detection in clutter," in *Conf. on Robot Learning*. PMLR, 2021, pp. 1602–1611.

[5] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, "Synergies between affordance and geometry: 6-dof grasp detection via implicit representations," *arXiv preprint arXiv:2104.01542*, 2021.

[6] 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 *Int'l. Conf. on Computer Vision*. IEEE/CVF, 2021, pp. 15 964–15 973.

[7] H. Huang, D. Wang, X. Zhu, R. Walters, and R. Platt, "Edge grasp network: A graph-based se (3)-invariant approach to grasp detection," *arXiv preprint arXiv:2211.00191*, 2022.

[8] J. Cai, J. Cen, H. Wang, and M. Y. Wang, "Real-time collision-free grasp pose detection with geometry-aware refinement using high-resolution volume," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1888–1895, 2022.

[9] J. R. Napier, "The prehensile movements of the human hand," *The Journal of bone and joint surgery. British volume*, vol. 38, no. 4, pp. 902–913, 1956.

[10] M. R. Cutkosky, "On grasp choice, grasp models, and the design of hands for manufacturing tasks." *IEEE Transactions on robotics and automation*, vol. 5, no. 3, pp. 269–279, 1989.

[11] X.-Y. Zhang, Y. Nakamura, K. Goda, and K. Yoshimoto, "Robustness of Power Grasp," in *Int'l Conf. on Robotics and Automation*. IEEE, 1994, pp. 2828–2835.

[12] Y. Zhang and W. A. Gruver, "Definition and Force Distribution of Power Grasps," in *Int'l Conf. on Robotics and Automation*, vol. 2. IEEE, 1995, pp. 1373–1378.

[13] B. Siciliano, O. Khatib, and T. Kröger, "Design of robot hands," in *handbook of robotics*. Springer, 2008, ch. 15.2, pp. 347–351.

[14] Robotiq, "2f-85 and 2f-140 grippers," 2024. [Online]. Available: https://robotiq.com/products/2f85-140-adaptive-robot-gripper

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

[16] C. Ferrari and J. F. Canny, "Planning optimal grasps," in *Int'l. Conf. on Robotics and Automation*, vol. 3, no. 4. IEEE, 1992, p. 6.

[17] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv preprint arXiv:1703.09312*, 2017.

[18] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Conf. on Computer Vision and Pattern Recognition*. IEEE/CVF, 2020, pp. 11 444–11 453.

[19] D. Winkelbauer, B. Bäuml, M. Humt, N. Thuerey, and R. Triebel, "A two-stage learning architecture that generates high-quality grasps for a multi-fingered hand," in *Proc of IEEE/RSJ Int'l. Conf. on Intelligent Robots and Systems*. IEEE, 2022, pp. 4757–4764.

[20] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, 2023.

[21] J. Weisz and P. K. Allen, "Pose Error Robust Grasping from Contact Wrench Space Metrics," in *Int'l Conf. on Robotics and Automation*. IEEE, 2012, pp. 557–562.

[22] K. Mirza and D. Grin, "General Formulation for Force Distribution in Power Grasp," in *Int'l Conf. on Robotics and Automation*. IEEE, 1994, pp. 880–887.

[23] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *Int'l. Conf. on Robotics and Automation*. IEEE, 2015, pp. 4304–4311.

[24] Y. Zhou and K. Hauser, "6dof grasp planning by optimizing a deep learning scoring function," in *RSS workshop on revisiting contact-turning a problem into a solution*, vol. 2, 2017, p. 6.

[25] C. Eppner, A. Mousavian, and D. Fox, "Acronym: A large-scale grasp dataset based on simulation," in *Int'l. Conf. on Robotics and Automation*. IEEE, 2021, pp. 6222–6227.

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

[27] A. Ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp Pose Detection in Point Clouds," *Int'l J. of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.

[28] R. Elandt, E. Drumwright, M. Sherman, and A. Ruina, "A pressure field model for fast, robust approximation of net contact force and moment between nominally rigid objects," in *Int'l. Conf. on Intelligent Robots and Systems*. IEEE/RSJ, 2019, pp. 8238–8245.

[29] R. Tedrake and the Drake Development Team, "Drake: Model-based design and verification for robotics," 2019. [Online]. Available: https://drake.mit.edu

[30] S. Haddadin, S. Parusel, L. Johannsmeier, S. Golz, S. Gabl, F. Walch, M. Sabaghian, C. Jähne, L. Hausperger, and S. Haddadin, "The franka emika robot: A reference platform for robotics research and education," *IEEE Robotics & Automation Magazine*, vol. 29, no. 2, pp. 46–64, 2022.

[31] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Conf. on Computer Vision and Pattern Recognition*. IEEE/CVF, 2019, pp. 5745–5753.

[32] 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 *Int'l. Conf. on Advanced Robotics*. IEEE, 2015, pp. 510–517.

[33] K. Shankar, M. Tjersland, J. Ma, K. Stone, and M. Bajracharya, "A learned stereo depth system for robotic manipulation in homes," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2305–2312, 2022.