Physics-Guided Hierarchical Reward Mechanism for Learning-Based Object Grasping

Yunsik Jung*, Lingfeng Tao*, Michael Bowman*, Jiucui Zhang^, and Xiaoli Zhang*, Member, IEEE

Abstract— Learning-based grasping can afford real-time motion planning of multi-fingered robotics hands thanks to its high computational efficiency. However, it needs to explore large search spaces during its learning process. The search space causes low learning efficiency, which has been the main barrier to its practical adoption. In addition, the generalizability of the trained policy is limited unless they are identical or similar to the trained objects. In this work, we develop a novel Physics-Guided Deep Reinforcement Learning with a Hierarchical Reward Mechanism to improve the learning efficiency and generalizability for learning-based autonomous grasping. Unlike conventional observation-based grasp learning, physics-informed metrics are utilized to convey correlations between features associated with hand structures and objects to improve learning efficiency and outcomes. Further, a hierarchical reward mechanism is developed to enable the robot to learn the grasping task in a prioritized way. It is validated in grasping tasks with a MICO robot arm in both simulation and physical experiments. The results show that our method outperformed the standard Deep Reinforcement learning method in task performance by 48% and learning efficiency by 40%.

I. INTRODUCTION

Multi-finger robotic grasping is critical to accomplish object manipulation that can replace human activities in various environments, such as the manufacturing industry, space, and deep-sea maintenance. Despite the potential, it is still challenging in several aspects. As high-dimensional robotic hands that can accomplish complex tasks have been developed, it significantly increases computational demands, which harms real-time performances in real-world applications. In addition, interactions between robotic hands and objects with various contours are demanding to accomplish stable performances. Therefore, the adoption of multi-finger robotic grasping is limited [1].

The Deep Reinforcement Learning (DRL) method has significantly improved performance by handling high-dimensional problems and enabling real-time autonomous grasping [5]. It provides an optimal control policy that maximizes an objective function/reward based on the specific state of an autonomous grasping task.

However, a common issue of DRL approaches and other observation-based robot learning is that training a robot is time-consuming to reach sufficient stabilities and performances. Training requires exploring a broad search space because of complex configurations of robotic hands and interactions with objects, which results in low learning efficiency. Further, the generalizability of the trained policy is limited unless they are identical or similar to the trained objects. One of the critical reasons for the learning efficiency and the generalizability issues is that current DRL methods mainly use task-related sparse reward components as the only criterion to define the reward function [6]. However, a successful grasp usually requires multiple quality-related criteria components, such as grasp pose, contact points/regions on the object, and grasp stability. These quality-related criteria are commonly used in physics-based grasping methods but have rarely been considered in learning-based grasping. Thus, it is not easy for current learning-based robots to fundamentally understand how to achieve and improve grasp quality other than task completion. Ideally, considering both task completion and these quality-related criteria as the reward can help RL-based robots to explore the environment and generalize the learned policy efficiently.

This paper introduces the physics-guided DRL with a hierarchical reward mechanism (PG-H-RL) for autonomous grasping and leverages both the positives of learning and physics to facilitate computationally efficient yet high-quality grasping solutions by enabling the robot to understand the problem at hand fundamentally. The contributions of this work are:

- Developed a physics-guided learning strategy for autonomous grasping, by integrating physics-based metrics as rewards to guide the robot to understand the grasping task. This improved learning efficiency and yielded physically consistent performances.
- Developed a hierarchical reward mechanism to learn the physics-based rewards in a prioritized logical way to help the robot to understand the priority of different metrics and improve learning efficiency.

II. RELATED WORK

A. Analytical Methods for Autonomous Grasping

Analytical approaches consider physics, kinematics, and dynamics of objects and hands to get the correct grasp, which is a vital aspect of accomplishing grasping tasks. In [10], they proposed a grasping simulator to compute robot and object motions under the influence of external forces and contacts. Form closure and force closure properties as basic grasp quality criteria were utilized in these approaches to find the
correct grasp [8]. Grasp quality measures have been developed to interpret the quality of robotic grasping. The measures associated with contact points on the object and the configurations of the robotic hands were developed [9, 10]. In [2], they introduced an approach combining empirical and analytical methods by imitating humans to reduce the computation time of calculating force-closure grasps. Finding the optimal solution to meet these criteria requires high computational power and limits real-world applications.

B. Learning-based Methods for Autonomous Grasping

DRL for robotic grasping has been actively studied in recent years. Earlier work aimed to acquire strategies for robotic grasps using DRL with images [3]. Recently, DRL methods have been proposed to accomplish autonomous grasping with vision-based observations [4]. However, these learning methods require extensive training data and time to explore. To overcome this, human preferences, demonstration data, and potential contact regions were utilized [11, 12, 13]. In [14], they estimated the probability of a successful grasp using the contact region database collected from human demonstrations. The probability was considered as a partial reward to increase learning speed. Although these pure learning approaches could improve learning efficiency by reducing the search space, they do not fundamentally learn physics as a physics-informed approach would.

C. Physics-guided Learning Strategies

Despite the validated effectiveness of physics-guided learning in other applications [15], few studies have been reported in the robotic grasping field. In [16], they used the physics-guided target poses as the input for the learning process to improve performance for manipulation tasks on a physics simulator. In [13], they proposed an RL method that utilized a grasp quality metric as the reward for a good grasping configuration by using the potential grasp locations estimated with the database of the contact information of successful grasps on the objects. In addition, [17] defined the reward function that summed the force-closure quality index [18]. However, all these methods treated the grasp quality as binary bonus rewards and used a linear summation which can be easily biased or lose the grasping information.

D. Hierarchical Reward in RL

The reward formulation of DRL is usually a linear summation of the reward components [6, 13, 14], which is implicit and inefficient for learning the multi-objective priorities and causes poor learning performance for multi-objective tasks (i.e., it takes a long time to learn or even fail to learn a correct policy). Hierarchical reward methods have been proposed to enable a robot to learn multi-objective tasks such as achieving autonomy or human-like merging actions for driving [19] and performing home service activities [20]. The formulation of the reward hierarchies contains logical or weighted connections. Logical connections are strict constraints, where the higher-level hierarchy must be learned before the lower-level one. Weighted connections are soft constraints, where the higher-level and lower-level hierarchies are learned together with a weighted summation. In [21], an RL agent for swarm robot control is trained with a logically connected hierarchical reward function. Inspired by these studies, this paper introduces the hierarchical reward in physics-guided grasp learning to learn multiple physics metrics and their correlations explicitly and efficiently.

II. METHODOLOGY

PG-H-RL is developed to enable a multi-finger robot to learn grasping and lifting an object to a target height, and perform additional practical tasks: firm grasp, insertion alignment, and transportation with motion disturbance. It is assumed that the object's position and contact information between the robotic hand and the object can be detected to calculate the grasp quality. The joints of the robotic arm and hand are controlled.

A. Reinforcement Learning Formulation

The grasping task is formulated as an RL problem that follows the Markov Decision Process (MDP). The MDP is defined as a tuple \( S, A, R, \gamma \), where \( S \) is the state of the environment and \( A \) is the set of actions. \( R(s, a) \) is the reward function to give the reward after the transition from state \( s \) to state \( s' \) with action \( a \) and \( \gamma \) is a discount factor. Since the task is required to consider interactions with objects and more accurate controls, the continuous control domain is considered. To solve the problem, PG-H-RL adopts the Twin Delayed Deep Deterministic policy gradient (TD3) algorithm, which is a model-free reinforcement learning [22].

B. Physics Metrics and Constraints

1) Contact points interacting with an object

Grasp quality is important for the agent to achieve a stable grasp. Contacts between the robotic hand and the object are important and necessary to evaluate grasp quality. The grasp matrix \( G \) is defined by the relevant velocity kinematics and force transmission properties of the contacts on convex/non-convex objects in three-dimensional space [10]. In this work, two measures of grasp quality are computed with \( G \) to evaluate the grasp quality: the measure for being graspable \( r_{\text{graspable}} \) and the normalized volume of the ellipsoid \( v_{\text{ew}} \). Inspired by previous studies with the traditional criteria of physics metrics as binary evaluations, the null space of the grasping matrix \( \mathbb{N}(G) \) is considered a reward to indicate whether a grasp is graspable or ungraspable based on internal object forces [10].

\[
    r_{\text{graspable}} = \begin{cases} 
    0 & \mathbb{N}(G) = 0 \\
    0.1 & \mathbb{N}(G) \neq 0 
    \end{cases} 
\]  \hspace{1cm} (1)

where \( \mathbb{N}(G) \neq 0 \) reveals being graspable. It judges the grasp quality by providing an initial guide before further evaluation, which can reduce the search space. The binary value is empirically determined considering its importance level relative to other components in the entire reward. Using \( G \), the contribution of all contact forces on the object can be detected to calculate the grasp quality. The joint of the robotic arm and hand are controlled.

The reward function that summed the force-closure quality index \( \mathbb{N}(G) \) is considered a reward to indicate whether a grasp is graspable or ungraspable based on internal object forces [10].

\[
    r_{\text{graspable}} = \begin{cases} 
    0 & \mathbb{N}(G) = 0 \\
    0.1 & \mathbb{N}(G) \neq 0 
    \end{cases} 
\]  \hspace{1cm} (1)

where \( \mathbb{N}(G) \neq 0 \) reveals being graspable. It judges the grasp quality by providing an initial guide before further evaluation, which can reduce the search space. The binary value is empirically determined considering its importance level relative to other components in the entire reward. Using \( G \), the contribution of all contact forces on the object can be detected to calculate the grasp quality. The joint of the robotic arm and hand are controlled.

\[
    Q_{\text{ew}} = \sqrt{\det(GG^T)} = \sigma_1 \sigma_2 \sigma_3 \cdots \sigma_m 
\]  \hspace{1cm} (2)

where, \( \sigma_1, \sigma_2, ..., \sigma_m \) denotes the singular values of \( G \), and \( m \) is the number of contact points on the object. This value is continuous and must be maximized to obtain the optimum grasp. In addition, the maximum value of \( Q_{\text{ew}} \) affected by the number of contacts is used to normalize the reward, \( r_{\text{ew}} \) as:

\[
    r_{\text{ew}} = \text{norm}(Q_{\text{ew}}) 
\]  \hspace{1cm} (3)

Fig. 1 (a) illustrates the examples of an optimal grasp of \( Q_{\text{ew}} \). An example of robotic grasps with their \( r_{\text{ew}} \)s is in Fig. 1 (b).
2) Hand pre-shaping constraints

Constraints were placed on the fingers to prevent closing the fingers and avoid contact between the fingers. These constraints act as penalties to the reward to pre-shape the robot hand for the following grasping tasks. The former provides a penalty when the fingers attempt to close before the hand approaches the pre-grasp distance, and the latter provides a penalty if there is any contact between the fingers.

C. Hierarchical Reward Mechanism with Physics Metrics

Using the physics metrics, multiple components in the reward function are prioritized logically to learn autonomous grasping progressively. An autonomous grasp task is broken down into three sequential stages: 1) approaching the object, 2) grasping the object, and 3) lifting the object from the table. In the approaching stage, the pre-shaping constraints were included in the reward function before touching and after grasping the object. The grasping stage includes grasping quality metrics with a hierarchical structure considering their priorities. The hierarchical structure reflects that a measure and/or constraint with a lower priority is not considered when a condition of one with a higher priority is not satisfied. It improves the learning efficiency because the agent can explore action/state spaces efficiently depending on the satisfaction of the higher level of hierarchies. The hierarchical physics-guided reward mechanism is shown in Fig. 2.

The reward function consists of multiple reward components for each stage. The approaching stage includes get close to the target (a penalty for the reward for the distance between the robotic hand and the object), prevent closing the fingers, and avoid contact between the fingers which are defined as penalties in Fig. 3. get close to the target can be defined as:

\[ p_{\text{target}} = - (v \times \text{dist}_{\text{obj hand}}) \] (4)

where \( v \) is a weighted coefficient and task-dependent as 10. In the grasping stage, pre-grasp preparation is a condition to determine whether the agent gets close enough to the object. The reward is the normalized exponential value of the distance between the robotic hand and the object, \( v_{\text{dist}} \):

\[ v_{\text{dist}} = \text{norm} (e^{0.1 \times \text{dist}_{\text{obj hand}}} \times d_{\text{grass}}) \] (5)

where \( \text{dist}_{\text{obj hand}} \) is the distance between the hand and the object. Being graspable is a condition to decide whether a grasp can grasp or not based on (1). If it reveals being graspable, then an additional reward, weight, is added to the reward, which is determined by prudent considerations to balance with other reward components.

Contact forces to grasp is a condition to show how much contact forces are applied to grasp the object based on \( v_{\text{rew}} \) in (3). In the lifting stage, \( r_{\text{obj height}} \) is calculated and added to the total reward, measuring the error between the current and target heights:

\[ r_{\text{obj height}} = \alpha \times \text{error}_{\text{obj height}} \] (6)

where \( \alpha \) is a weighted coefficient and \( \beta \) is the maximum error of \( \text{error}_{\text{obj height}} \). Fig. 3 shows the total reward function. \( \lambda, \mu, \) and \( v \) are binary coefficients defined as 1 when both conditions of higher hierarchies and the corresponding condition are satisfied; otherwise, they are 0. They can be described as:

\[ \lambda = \begin{cases} 0 & \text{if dist}_{\text{obj hand}} \geq d_{\text{grass}} \\ 1 & \text{if dist}_{\text{obj hand}} < d_{\text{grass}} \end{cases} \] (7)

\[ \mu = \begin{cases} \lambda \times 1 & \text{if } Q_{\text{rew}} = 0 \\ 0 & \text{if } Q_{\text{rew}} \neq 0 \end{cases} \] (8)

\[ v = \begin{cases} \lambda \times \mu \times 1 & \text{if } Q_{\text{rew}} > 0 \\ 0 & \text{otherwise} \end{cases} \] (9)

where \( d_{\text{grass}} \) is a boundary entering the grasping stage.

III. EXPERIMENTS

A. Experimental Setup

1) MICO arm: To accomplish the task, a Kinova’s MICO arm is used [23], which has 6 rotational joints for the arm and the three-fingered robot hand. The MICO arm can be controlled to open and close the fingers to grasp the object.

2) Simulator: CoppeliaSim (V-REP) is used as the simulator [24]. It provides a precise physics engine for interactions between the environment, the robot, and the object. Scripts were programmed in the scene using V-REP API to connect the kinematics and the sensing details. The maximum number of steps for each training episode is 300.

For the TD3 agent, we define the state observation, including the object’s positions, the finger joint angles, the
The success of the orientation less than 45º, and the success of the transportation reorientation task is determined within an error of the object's height. The success of the firm grasp task is determined when the object is lifted within 0.02 m error to the target (Fig. 6 (a)). The success of the grasping task is determined between the desired location of the object center and the object's actual height and the target of 0.05 m from the table. The results with the cube are relatively higher than other shapes. The polyhedron results are the worst due to the bigger variances due to overfitting or overwriting the agent's learning. The above results validate that considering the physics metrics and constraints as the reward effectively guides agents to learn the task faster. However, introducing more objectives without considering their relative priorities confuses the agent to balance among different reward components and makes the learning unstable.

### B. Learning Efficiency

The experiments executed 5 cases, meaning agents learned the task 5 times for each method to generate statistical results, including the means and variances during training. The total reward indicates the earned rewards during 300 steps for an episode, and Fig. 5 shows the means of the total rewards for each training episode. The results reveal that PG-H-RL reaches 56.40% of its maximum total reward with 200 training episodes, while Task Only reaches 34.04% and Linear Summed reaches 47.25%. Further, each method’s convergence rate was estimated from 400 to 600 training episodes with a standard deviation. PG-H-RL’s rate was 0.01, Task Only’s rate was 0.04, and Linear Summed’s rate was 0.06. PG-H-RL shows a steadier trend with the mean of the total rewards, while the other two methods show greater variances due to overfitting or overwriting the agent’s experiment by the new experience. The above results validate that considering the physics metrics and constraints as the reward effectively guides agents to learn the task faster.

### C. Learning Outcome

Although the success rates for the methods indicate there are still failure cases due to the difficulties in predicting the interactions between the robotic hand and the object or the environment with the multi-finger robotic system, PG-H-RL outperforms Task Only and Linear Summed with various numbers of training episodes, 200, 600, and 1000. Table II shows the success rates for different shapes and sizes of the object performed with the learned policies with 1000 training episodes. Since the policies are trained with the cube, the results with the cube are relatively higher than other shapes. The polyhedron results are the worst due to the bigger differences in the contact positions. The size reduction of 10%

### IV. RESULTS AND DISCUSSION

#### A. Hierarchical Reward Mechanism

Fig. 4 shows an episode reward using PG-H-RL in late training. The duration from 0 to 21 steps indicates the robotic hand is approaching the object. Then, the reward entered the next stage for grasping with three physics-guided reward components associated with the object’s perspective. The pre-grasp preparation condition is the first hierarchy and is satisfied after 21 steps. The being graspable condition is the second hierarchy and is satisfied after 25 steps. The contact forces to grasp condition is the third hierarchy and is satisfied after 29 steps. The hierarchical reward mechanism allows the agent to learn higher level constraints before lower ones.

#### B. Learning Efficiency

The experiments executed 5 cases, meaning agents learned the task 5 times for each method to generate statistical results, including the means and variances during training. The total reward indicates the earned rewards during 300 steps for an episode, and Fig. 5 shows the means of the total rewards for each training episode. The results reveal that PG-H-RL reaches 56.40% of its maximum total reward with 200 training episodes, while Task Only reaches 34.04% and Linear Summed reaches 47.25%. Further, each method’s convergence rate was estimated from 400 to 600 training episodes with a standard deviation. PG-H-RL’s rate was 0.01, Task Only’s rate was 0.04, and Linear Summed’s rate was 0.06. PG-H-RL shows a steadier trend with the mean of the total rewards, while the other two methods show greater variances due to overfitting or overwriting the agent’s experiment by the new experience. The above results validate that considering the physics metrics and constraints as the reward effectively guides agents to learn the task faster. However, introducing more objectives without considering their relative priorities confuses the agent to balance among different reward components and makes the learning unstable.

#### C. Learning Outcome

Although the success rates for the methods indicate there are still failure cases due to the difficulties in predicting the interactions between the robotic hand and the object or the environment with the multi-finger robotic system, PG-H-RL outperforms Task Only and Linear Summed with various numbers of training episodes, 200, 600, and 1000. Table II shows the success rates for different shapes and sizes of the object performed with the learned policies with 1000 training episodes. Since the policies are trained with the cube, the results with the cube are relatively higher than other shapes. The polyhedron results are the worst due to the bigger differences in the contact positions. The size reduction of 10%

### TABLE I. HYPER-PARAMETERS FOR TD3 ALGORITHM

| Parameters       | Values          |
|------------------|-----------------|
| Sample Time      | 0.05 [sec]      |
| Discount Factor  | 0.99            |
| Mini-Batch Size  | 256             |
| Experience Buffer Length | 1e+6   |
| Target Smooth Factor | 0.005  |
| Learning rate    | 0.001           |
| Target Update Frequency | 2     |
| Sequence Length  | 1              |

The robotic hand’s position, and the arm’s joint angles. Table I shows the hyper-parameters for the TD3 agent. α and β in (6) are empirically determined as 30 and 0.05. In training, the cube with a side length of 0.065 m was used to train the policy. The cylinder and polyhedron and the various sizes as ±10% of the original size of objects were used to evaluate the generalizability to different objects since they are similar to the trained object but have different contours and require different contact points and grasping shapes. The trained policy performs the grasping task that grasps and lifts the object to a target height to evaluate the basic grasping performance. Further, the firm grasp, insertion alignment, and transportation with motion disturbance tasks are designed since they are common and critical to evaluate the grasping stability of the trained policy. Unstable grasps can easily lead task failures by dropping or sliding the object while rotating or transferring the object, which indicates that their successes are highly sensitive to grasp quality.

### B. Evaluation Methods and Metrics

Two different reward functions are considered baselines to evaluate the influence of the physics metrics and the hierarchical reward mechanism of the PG-H-RL method. Except for the reward functions, the baselines are in the same conditions and set up with PG-H-RL. A baseline, Task Only, has a reward function with task-related reward components:

\[ r_{\text{Task Only}} = P_{\text{target}} + r_{\text{obj.height}} \quad (10) \]

It uses a linear summation instead of a hierarchical reward mechanism. The second baseline, Linear Summed, uses all reward components considered in PG-H-RL, but a linear summation is used instead of a hierarchical reward mechanism. The PG-H-RL method and the two baselines were trained with the same grasping task to evaluate the learning efficiencies and outcomes. Trends of the total reward at different training episodes are used to compare the learning efficiency. The trends are evaluated by comparing how fast the total reward increases and maintaining this increase. To evaluate the learning outcomes, an error of height between the object’s actual height and the target of 0.05 m from the table is used. \( Q_{\text{vew}} \) is considered to assess the stable and firm grasp quality. The distance to the object center is the distance between the desired location of the object center and the actual object center to evaluate the stability of a grasp pose (in Fig. 6 (a)). The success of the grasping task is determined when the object is lifted within 0.02 m error to the target height. The success of the firm grasp task is determined when \( Q_{\text{vew}} \) > 5 and \( \text{dis}_{\text{obj.hand}} \) < 0.1 m. The success of the reorientation task is determined within an error of the object orientation less than 45º, and the success of the transportation is determined when the object is relocated in the target boundary with a radius of 0.09 m.

### Figure. 4 The reward evolution for a single episode using PG-H-RL.

![Figure 4](image-url)
reduced success rates for all methods and shapes since the inherent difficulty of grasping a small object relative to the large robot hand. PG-H-RL consistently outperforms Task Only and Linear Summed for both shapes and sizes since PG-H-RL considers the physics metrics, making it more generalizable to different objects. The results with 200 and 600 numbers of training episodes also have consistent inclinations in the performances. It is difficult to reach a 100% success rate since the training environment is stochastic, with uncertainty and noise that may cause task failure. DRL brings stochastic behavior and leads to different grasping behaviors. Especially for the grasping task, the failures occur for several reasons such as wrong approaching direction, finger closing timing, and interactions with the object. It indicates that learning still cannot fully cover all possible testing cases. All methods are influenced the same, but PG-H-RL outperforms the baseline methods due to utilizing the physics metrics and the hierarchical reward mechanism. To confirm statistically significant differences for the success rates between PG-H-RL and the baselines, N-1 Chi-Square test [25] that compares two binary variables for two independent groups is used to calculate p-values in Table III. The p-value of the comparison between PG-H-RL and Task Only with 1000 training episodes are not statistically significant, while all other p-values are statistically significant (at a p-value level of 0.05).

**Table II. Results of Success Rates for Various Shapes**

| Shape     | Size | PG-H-RL [%] | Task Only [%] | Linear Summed [%] |
|-----------|------|-------------|---------------|-------------------|
| Cube      | -10% | 72          | 66            | 44                |
|           | Original | 86          | 72            | 56                |
|           | +10%   | 90          | 82            | 60                |
| Cylinder  | -10% | 50          | 32            | 40                |
|           | Original | 78          | 56            | 44                |
|           | +10%   | 80          | 54            | 48                |
| Polyhedron| -10% | 4           | 1             | 2                 |
|           | Original | 42          | 16            | 28                |
|           | +10%   | 64          | 36            | 30                |

**Table III. P-Values Using One-Way Analysis of Variance for Learning Performances**

| Category                | PG-H-RL vs Task Only | PG-H-RL vs Linear Summed |
|-------------------------|----------------------|--------------------------|
| Error of Height         | 0.4474               | 8.9922e-18               |
| Distance to the Object Center | 9.8952e-16           | 5.8884e-13               |
| Grasp Quality           | 1.5815e-15           | 4.0366e-19               |

**Table VI. P-Values of N-1 Chi-Square Test for Success Rates**

| Training episodes | PG-H-RL vs Task Only | PG-H-RL vs Linear Summed |
|-------------------|----------------------|--------------------------|
| 200               | 0.0074               | 1.3912e-5                |
| 600               | 0.0004               | 0.0014                   |
| 1000              | 0.2183               | 8.6378e-5                |

a. The One-way Analysis of variance for the performance data can be obtained using the MATLAB function (ANOVA1). The p-value from the distributions can be derived by the function which performs the One-way Analysis of Variance.

b. The N-1 Chi-squared distribution (p-value) can be obtained using the MATLAB function (CHIDIST).

The mean and variance of the normalized total rewards for each method. The shaded areas indicate the standard deviations for the 5 cases.
performances in the error of height. Given that the Task Only reward solely focuses on task-related rewards with a specific lifting height, this similarity shows that the PG-H-RL can learn the additional grasp-quality-related rewards without sacrificing the conventional task-related reward. Further, Table V shows the experimental results for the practical tasks: firm grasp, insertion alignment, and transportation with motion disturbance. The success rates of all three methods drop significantly from the basic grasp task to the insertion and transportation tasks due to the additional object pose requirements and/or motion disturbance. PG-H-RL still outperforms the baseline methods over all the tasks and shapes, even the results of PH-H-RL also show declining performances for the untrained shapes for all the tasks.

Fig. 7 illustrates the performance trained using PG-H-RL and the performance trained using Task Only for a specific grasp instance when the object reaches the target height. Although both cases succeed in the task, the former result shows a higher grasp quality with three contacts to secure the object more stably. In contrast, the latter result shows a lower grasp quality with two contacts with the object. Since Task Only encourages task related rewards but cannot provide finer reward differences in grasp quality, it is difficult to achieve secure grasps. In contrast, the PG-H-RL method can provide a finer evaluation of postures since it considers task-related rewards and grasping quality.

Our results show PG-H-RL can not only learn autonomous grasping skills but also improve performances and learning efficiency by using physics metrics and hierarchical reward mechanism.

**REFERENCES**

[1] F. Ficuciello, G. Palli, C. Melchiorri and B. Siciliano, "A model-based strategy for mapping human grasps to robotic hands using synergies," IEEE/ASME International Conference on Advanced Intelligent Mechatronics, pp. 1737-1742, 2013

[2] El-Khoury, S., & Sabbani, A., "A new strategy combining empirical and analytical approaches for grasping unknown 3D objects," Robotics and Autonomous Systems, 58(5), 497–507, 2010

[3] Moussa, et al., “Connectionist model for learning robotic grasps using reinforcement learning,” IEEE International Conference on Neural Networks, 3, 1771–1776, 1996

[4] Kalashnikov, D., Ir pan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V., & Levine, S., “Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation,” ArXiv, CorRL, 1–23, 2018

[5] I. Popov, et. al. “Data-efficient Deep Reinforcement Learning for Dexterous Manipulation.” ArXiv abs/1704.03073, 2017

[6] Rajeswaran, A., et al., “Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations,” arXiv:1709.10087, 2018.

[7] A. T. Miller, P. K. Allen, "Graspit! A versatile simulator for robotic grasping." IEEE Robotics & Automation Magazine, 11(4), pp. 110-122, 2004

[8] E. Rimon and J. Burdick, “On force and form closure for multiple finger grasps,” Proc. IEEE ICRA, pp. 1795–1800, 1996

[9] Roa, M. A., & Suárez, R., “Grasp quality measures: review and performance,” Autonomous Robots, 38(1), pp 65–88, 2014

[10] Prachtichzzo, D., & Trinkle, J. C., “Springer Handbook of Robotics,” In Springer Handbook of Robotics, 2008

[11] Pinsler, R., Akrou, R., Osa, T., Peters, J., & Neumann, G., “Sample and Feedback Efficient Hierarchical Reinforcement Learning from Human Preferences,” IEEE International Conference on Robotics and Automation (ICRA), pp. 596–601, 2018

[12] Mandikal, P., & Grauman, K., “Dexterous Robotic Grasping with Object-Centric Visual Affordances,” 1–11, 2020, http://arxiv.org/abs/2004.04329

[13] Osa, T., Peters, J., & Neumann, G., “Hierarchical reinforcement learning of multiple grasping strategies with human instructions,” Advanced Robotics, 32(18), 955–968, 2018

[14] E. Valarezo Añazco et al., “Natural object manipulation using anthropomorphic robotic hand through deep reinforcement learning and deep grasping probability network,” Appl. Intell., 2020

[15] Zhao, P., & Liu, Y., “Physics Informed Deep Reinforcement Learning for Aircraft Conflict Resolution,” IEEE Transactions on Intelligent Transportation Systems, 1–14, 2021

[16] Garcia-Hernando, G., Johns, E., & Kim, T. K., “Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning,” IEEE International Conference on Intelligent Robots and Systems (IROS), 9561–9568, 2020

[17] Monforte, M., & Ficuciello, F., “A Reinforcement Learning Method Using Multifunctional Virtual Component Analysis for Human-like Grasping,” IEEE Transactions on Cognitive and Developmental Systems, 8920(c), 1–1, 2020

[18] Bicchi A. “On the closure properties of robotic grasping. International Journal of Robotics Research,” 14:319–334, 1994

[19] Sun, L., “Intelligent and High-Performance Behavior Design of Autonomous Systems via Learning, Optimization and Control,” Ph. D Thesis, Univ. of California, Berkeley, California, 2019.

[20] Zhang, M., et al. “Service skill improvement for home robots: Autonomous generation of action sequence based on reinforcement learning,” Knowledge-Based Systems, 212, 106605, 2021

[21] Clayton, N. R., & Abbass, H., “Machine Teaching in Hierarchical Genetic Reinforcement Learning: Curriculum Design of Reward Functions for Swarm Shepherding,” 2019 IEEE Congress on Evolutionary Computation, CEC 2019 - Proceedings, 1259–1266, 2019

[22] Fujimoto, S., Van Hoof, H., & Meger, D., “Addressing Function Approximation Error in Actor-Critic Methods,” 35th International Conference on Machine Learning, ICML 2018, 4, 2587–2601, 2018

[23] Robotics company—Robotic assistive technology—Kinova. https://www.kinovarobots.com/en. Accessed 18 Feb 2021

[24] E. Rohmer et al., “V-REP: A versatile and scalable robot simulation framework,” IEEE IROS, pp. 321–326, 2013.

[25] Sauro, J., Lewis, J., “Quantifying the user experience,” 2nd ed., Netherlands: Elevier Inc. of Publisher, ch. 5, pp. 61-102, 2016

[26] Wu, C. F. J., Hamada, M. S., “Experiments: Planning, Analysis, and Parameter Design Optimization,” 2nd ed., Wiley of Publisher, 2009

**TABLE V. SUCCESS RATES FOR PRACTICAL TASKS [%]**

| Shape       | Cube | Cylinder | Polyhedron |
|-------------|------|----------|------------|
| PG-H-RL     | 79   | 48       | 64         |
| Task Only   | 50   | 25       | 35         |
| Linear Summed | 30 | 23       | 20         |

a. The tasks: task 1 is the firm grasp, task 2 is the insertion alignment, and task 3 is the transportation with motion disturbance.

b. All results are percentages of success rates.