

HERB: Human-augmented Efficient Reinforcement learning for Bin-packing

Gojko Perovic¹ Nuno Ferreira Duarte² Atabak Dehban² Gonçalo Teixeira²
Egidio Falotico¹ José Santos-Victor²

Abstract— Packing objects efficiently is a fundamental problem in logistics, warehouse automation, and robotics. While traditional packing solutions focus on geometric optimization, packing irregular, 3D objects presents significant challenges due to variations in shape and stability. Reinforcement Learning (RL) has gained popularity in robotic packing tasks, but training purely from simulation can be inefficient and computationally expensive. In this work, we propose HERB, a human-augmented RL framework for packing irregular objects. We first leverage human demonstrations to learn the best sequence of objects to pack, incorporating latent factors such as space optimization, stability, and object relationships that are difficult to model explicitly. Next, we train a placement algorithm that uses visual information to determine the optimal object positioning inside a packing container. Our approach is validated through extensive performance evaluations, analyzing both packing efficiency and latency. Finally, we demonstrate the real-world feasibility of our method on a robotic system. Experimental results show that our method outperforms geometric and purely RL-based approaches by leveraging human intuition, improving both packing robustness and adaptability. This work highlights the potential of combining human expertise-driven RL to tackle complex real-world packing challenges in robotic systems.

I. INTRODUCTION

Irregular object packing represents an important skill for robotic applications in multiple industrial or household domains. However, even the two-dimensional discrete bin-packing problem, which only considers regular objects, has been proven to be NP-hard [1]. Despite its challenging nature, due to the prominence of this task, many solutions and task considerations have been proposed in the literature [2], [3], [4]. By considering objects with general irregular shapes, higher demands are placed on the capabilities of the robotic system [3]. Notably, recent data-driven methods are commonly used to address irregular object packing [5], [6], [7].

Humans can pack multiple objects in a box while simultaneously considering multiple constraints such as affordances of which object can be placed on top of the other. Thus, while the human performance can provide insight in solving

Corresponding author - GP: g.perovic@santannapisa.it

¹BioRobotics Institute, Scuola Superiore Sant'Anna, Pisa, Italy and with the Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Pisa, Italy. {gojko.perovic, egidio.falotico}@santannapisa.it

²Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal. {nferreira@duarte, adehban, jasv}@isr.tecnico.ulisboa.pt

This work was supported by the LARSyS FCT funding - 10.54499/UDB/50009/2020, the Lisbon ELLIS unit, the Center for Responsible AI and MPR-2023-12- SACCCT- Project 14935 AI.PackBot.



Fig. 1. Baxter® robot with the box and objects from the BoxED dataset. The heightmap images capturing the state of the box are taken with a depth camera (out of crop).

geometrical problems, it can also be a source of latent information on object properties (e.g., fragility, deformability, compatibility) that would otherwise be unavailable to a given robotic system. This information could be used to alleviate constraints on achieving a successful task completion while providing a qualitatively better result.

Furthermore, learning from humans has been at the forefront of recent advances in robotic manipulation [8]. However, collecting data from humans is cumbersome and time-consuming. We envision a learning scenario where the discrete aspects of the task which have a smaller search space (in this case, the sequence) are learned from human data whereas the continuous, high dimensional actions (exact placing location and orientation inside the box) are learned via Reinforcement Learning (RL).

With these reasons in mind, this paper presents three main contributions:

- We propose a Human-augmented Reinforcement learning-based Bin-packing (HERB) method endowed with human-like sequence prediction algorithm to solve irregular object packing;
- We set up a simple RL environment to efficiently train and evaluate the performance of the proposed method;
- To validate our approach and to qualitatively assess packing, a robotic system including a Baxter® robot [9] is developed (Fig. 1).

The rest of this paper is organised as follows: In Sec. II an overview of the literature with the focus on packing 3D irregular objects and robotic solutions is given. The proposed approach is outlined in Sec. III. Experimental Protocol,

including the Dataset, Environment, Evaluation Procedure, and Metrics, are specified in Sec. IV. The Model Selection, Task Performance Metrics, and Qualitative Metrics results are presented in Sec. V. Then, we validate the feasibility of the proposed model in a real-world setting in Sec. VI. Finally, our findings are summarized and discussed in Sec. VII, and insight into the potential for further work is outlined in Sec. VIII.

II. RELATED WORK

Object packing has been a long-standing challenge due to its high practical significance and inherent complexities [2], [3], [4]. In the most common definition, the planning segment of the problem is commonly referred to as the Bin Packing Problem [4], where sets of cuboid objects must be placed in discrete bins of a larger container. As we aim to contextualize object packing as a skill for a robotic manipulator, we consider irregular 3D objects and a continuous placement space.

Under these considerations, closed-form solutions for irregular object packing are not available. Thus, in 3D, irregular object packing planning methods can largely be divided into *Heuristic* or *RL-based* methods.

Heuristic solutions aim to solve the packing by employing geometrical estimates of the optimal object placement. While these methods can be well-performing, they tend to be computationally expensive in inference [10], [5]. Furthermore, by defining a heuristic, the solution space could be more constrained, which might lead to difficult transfer to real-world scenarios. For example, exact 3D models of the objects might be missing, or the perception framework might not be precise enough to observe the voxel occupancy of the box. More critically, if practical scenarios are considered, these methods tend to strictly focus on solving the combinatorial problem and space optimization to the detriment of the physical feasibility of the solution [5].

RL-based [11] solutions aim to use a data-driven approach to solve the planning problem [12], [6], [7]. In the case of regular cuboid objects, and discrete bins, straightforward RL solutions tend to work well [12]. When irregular objects are considered, the algorithm design becomes significantly more complex, leading to larger models [6], [7]. Thus, RL-based methods rely on either discretizing the action space by binning [6] or employing a geometric-based heuristic [7] and then running an exhaustive search to estimate the state-action value Q of candidate solutions. While such discretizations simplify the RL approach, they constrain the potential generalization of the method, and in many cases the exhaustive searches lead to long training times [6], [7].

Vision-based systems are sometimes unable to perceive all pertinent object information. Recent work by Chen et al. [13] proposed integrating visual and tactile information to reorder the packing sequence based on deformation, placing the deformable objects last. In the proposed work, we aim to exploit the information embedded in human decisions to similarly augment the packing sequence based on the implicit object properties. Going beyond purely geometrical,

vision-based features is crucial in enabling novel packing approaches with advanced capabilities.

Learning from humans has emerged as one of the leading paradigms when considering data-driven approaches for acquiring complex robotic skills [8]. In previous work, we proposed a dataset referred to as *Box packing with Everyday items Dataset (BoxED)* [14]. Furthermore, we proposed a data-driven algorithm for human-like packing sequence generation [15]. Here, we apply this algorithm to alleviate the training and implementation complexity of the RL algorithm and generate object packs with human-like qualities.

III. PACKING ALGORITHM

The packing algorithm consists of two submodules, *Human-like Sequence Planning* and *Placement Prediction*.

A. Human-like Sequence Planning

Given a set of available objects $O = \{obj_A, obj_B, \dots\}$, and the dataset of human packing sequences [14], a model of human-like transitions can be modeled. More specifically, a static Markov chain based on pairwise transitions $obj_A \rightarrow obj_B$ is reconstructed [15]. By applying a modifying beam-search algorithm, the obtained transition matrix can be sampled to generate human-like sequences. More specifically, experiments have shown that constraining the modified beam search to sample the next 3 objects (referred to as Beam-3) produces the most human-compatible sequences [15].

Thus, given a list of objects, we generate a human-like sequence using the Beam-3 method. The ordered sequence should not only be beneficial in terms of size and geometry compatibility, but importantly, implicit information in human sequence preferences can be exploited.

B. Placement Prediction

Soft Actor Critic (SAC) is employed as the backbone RL algorithm [16] for placement prediction learning. A solid algorithm backbone is integral to focus on task-specific hyperparameter settings (such as sequence planning and reward function) while keeping the algorithm hyperparameters fixed. Thus, SAC is chosen due to its robustness to hyperparameter settings, its ability to perform continuous predictions, and relatively fast convergence (compared to other policy gradient methods).

1) *Observation and Action spaces*: An image of a box heightmap is used as the basis for observation of the RL algorithm [6], [7]. A top-down projection of the next object to be packed is appended to the box heightmap, and then the complete image is padded to 224×224 pixels (as seen in Fig. 2).

The policy output is a three-dimensional vector representing the object's x and y position and the planar rotation θ . When predicting the pose, the vertical coordinate z can be estimated from the box heightmap and the object model [6], [7], thus reducing the action space. The motivation behind constraining the rotation to be planar is two-fold. Firstly, it is challenging to re-orient objects in a practical robotic scenario, especially while minding possible collisions [2],

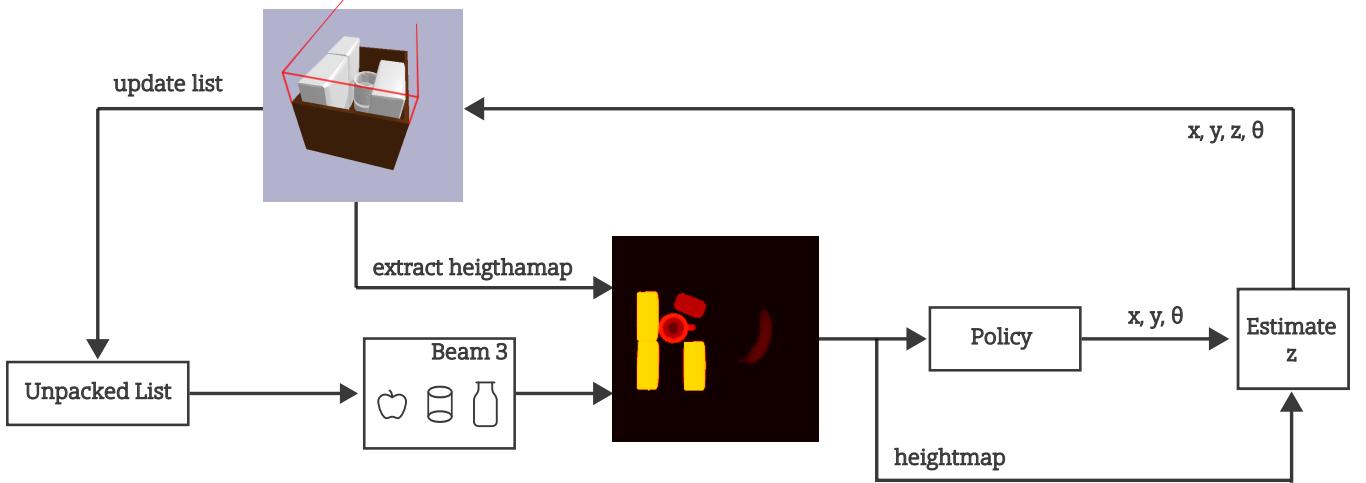


Fig. 2. Complete block diagram of the proposed system in inference. Given a list of unpacked objects, the Beam-3 algorithm sorts the candidates. Then, the projection of the object is concatenated to the state of the box represented by a heightmap. Based on this, the policy predicts the x, y, θ , which are then used to estimate the vertical position z . The episode terminates on the successful placement of all the objects, or should an object overcome the sides of the box or the vertical constraint (denoted by red lines).

[7]. Secondly, by increasing the action space dimension, the learning problem would become increasingly complex, potentially without performance increase at a given task.

2) *Reward Function*: We consider three reward functions. The *Simple* reward denotes a reward of 1 if the object is placed inside the box, and -1 otherwise. For the second reward modality, *Compactness* (sometimes referred to as *utility*) of packing is considered. While it can be defined differently across literature [10], [6], [7], we adopt a definition akin to [6]:

$$C = \frac{V_{cml}}{V_{min_box}} \quad (1)$$

where V_{cml} is the cumulative volume of objects placed inside the box, and $V_{box,min}$ is the volume of the minimal box encapsulating them. This definition supports the dataset [14], and thus human placements since it is expected that the taller objects will overcome the height of the box. Thus, the agent is rewarded C if the object is placed inside the box, and -1 otherwise. As for the third reward function, a combination of *Compactness* and *Stability* (CS) is evaluated. We define *Stability* as the violation of rotation constraint (i.e., the object tipped over the threshold of 10°). Thus, the agent obtains a reward of 1 if the rotation constraint was not violated and 0 otherwise. The CS reward is then composed as:

$$CS = \begin{cases} \alpha C + (1 - \alpha)S & \text{if object_inside} \\ -1 & \text{else} \end{cases} \quad (2)$$

where hyperparameter $0 \leq \alpha \leq 1$ is used to trade off compactness and stability.

A complete block diagram of the proposed approach is shown in Fig. 2.

IV. EXPERIMENT

A. Dataset and Environment

The environment for learning and evaluation is based on [15], however, while the original environment was im-

plemented in Unity [17] physics engine, we decide to reimplement it in PyBullet [18]. This is done to better exploit parallel simulations, facilitating faster RL training. To enable the reproduction of human placement from BoxED [14], appropriate transforms are applied. To further enhance the simulation speed, meshes from the dataset are simplified to be convex and watertight [19], [7]. The simulation is wrapped as a Gymnasium environment [20].

When framed within the RL context, each step represents one object to be packed, and each episode represents one experimental session defined by BoxED experiment [15]. The episode is terminated on successful placement of all the available objects or on an unsuccessful (e.g., out of the box) placement of any object. Since the height of the original box (16.4 cm) is shorter than certain dataset objects, a vertical margin of 13 cm is added to the successful placement bounding box (Fig. 2).

B. Evaluation Procedure and Metrics

We employ BoxED [14] trials and packing sequences as the test episodes for our experiments. We evaluate the performance of different packing approaches and compare them with human packs considering the following objective task performance metrics: *Success rate*, which is defined as the percentage of successful test episodes, *Number of packed objects*, which represents a distribution of packed objects over test episodes, and *Latency*, the single-step inference time distribution.

We also report *Compactness* and *Stability* metrics across the test episodes. While these metrics are qualitative, they provide significant insight into the performance of different algorithms.

V. RESULTS

A. Model Selection

We study the impact of several design choices by running different flavors of HERB. These variations include the importance of selecting a random sequence of objects vs. Beam-3, and different reward functions introduced in Sec. III-B. All models were trained on an *Ubuntu 22.04.5* machine with *Intel Xeon Max 9460* CPU and *NVIDIA A30 PCIe* GPU. On this computer, training one seed for 10M steps takes ~ 9 hours.

Fig. 3 reports the Normalized Reward and Episode Length (equivalent to the number of objects packed) across different settings. The reward is Min-Max normalized across three parameter-setting seeds run for each configuration.

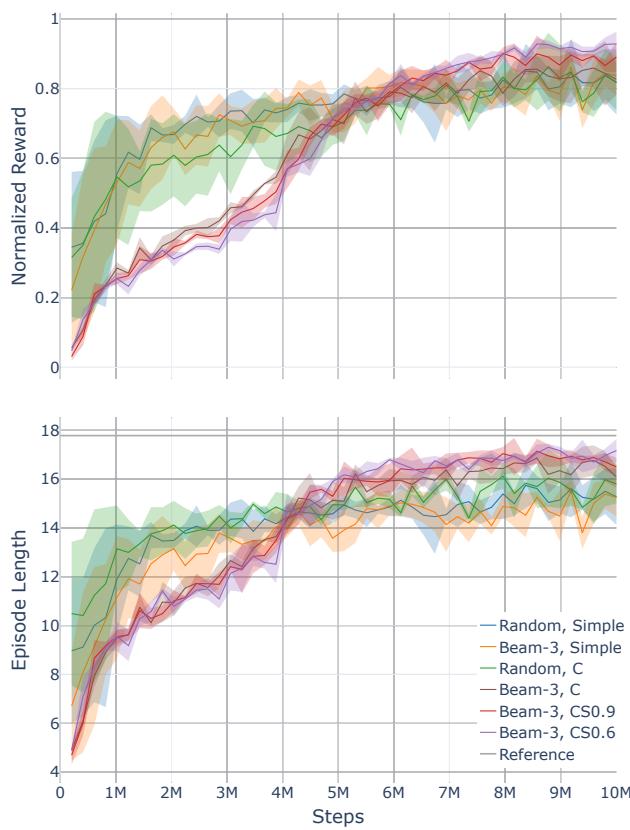


Fig. 3. Normalized Reward and Episode Length mean and standard deviation over three seeds for each setting. Higher is better. Reference in Episode Length plot denotes the mean number of objects to pack, set forth by the RL environment.

As the two best-performing parameter settings, CS0.9 and CS0.6, display comparable performance in terms of RL metrics, they will be further analyzed in the following sections.

B. Task Performance and Qualitative Metrics

In this subsection, tests are performed on packing sequences from the BoxED [14] and compared to human performance. Table V-B reports the task performance metrics

across different methods and human packing. As a baseline comparison, PackIt Heuristic [5] is used, which consists of sequence planning by the largest object, rotation alignment, and Bottom-Left-Back Fill (BLBF) position selection. We consider two implementations of PackIt Heuristic, the SO(2) which limits the orientation to planar rotations, and the SO(3) [5]. For human packing, BoxED trials are replayed in the RL environment. To evaluate latency, the execution time of the placement prediction of the complete pose is considered. Concerning the HERB algorithm, latency includes the Policy and Estimate z submodules (Fig. 2). PackIt Heuristic SO(2) latency only includes the BLBF module (since the objects are already rotation aligned). Finally, PackIt Heuristic SO(3) latency includes the rotation alignment and BLBF modules. The inference is run on a *Ubuntu 20.04.6* machine with *Intel i7-900* CPU and *NVIDIA GeForce RTX 2060* GPU.

TABLE I
TASK PERFORMANCE METRICS

	Success Rate [%]	No. Objects mean (std)	Latency [ms] mean (std)
Reference	100	17.40 (3.63)	- -
BoxED [14]	84.41	15.22 (4.79)	- -
PackIt SO(2) [5]	82.50	16.33 (4.46)	1.59 (0.28)
PackIt SO(3) [5]	55.89	13.92 (5.07)	2.51 (0.52)
HERB CS0.9	87.83	16.76 (3.79)	0.83 (0.26)
HERB CS0.6	86.31	16.43 (4.26)	0.77 (0.24)

Fig. 4 shows the compactness and stability distributions. To estimate stability of the BoxED placements and PackIt Heuristic SO(3), the threshold of 10° is considered between the intended pose of the object, and the resulting one.

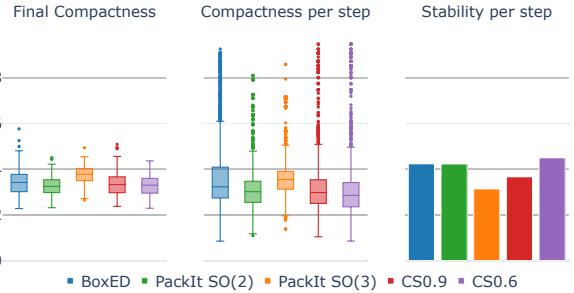


Fig. 4. Final Compactness of the box, Compactness per step, and Stability per step for different approaches on the BoxED sequences.

VI. ROBOTIC SYSTEM

To validate the feasibility of the proposed method, a packing robotic system is set up (Fig. 1).

To obtain the box heightmap *Intel ® RealSense™ L515* LiDAR camera is used. The image obtained from the sensors is cropped, padded, and threshold-filtered to mitigate noise around the bottom of the box. Two best performing models (HERB CS0.9 and HERB CS0.6) are qualitatively assessed to determine their robustness to real-world input and predicted poses. A demonstration of this assessment is performed by perturbing the predicted poses under real

sensory input (Fig. 5). Following this assessment, we chose CS0.6 as the preferable parameter setting for operation. We further discuss our findings in Sec. VII.

A Baxter ® bi-manual robot equipped with two electric parallel grippers [9] is selected as our robotic platform. The range of one gripper is adjusted for grasping smaller objects, and that of the other is adjusted to grasp larger objects from the dataset. We used *ROS Noetic* [21] to facilitate picking, placing, and interfacing with HERB. A video demonstrating the HERB system performance is included as supplementary material.

VII. DISCUSSION

A. Experimental Results

From the ablation studies (Fig. 3), it can be concluded that both the C and CS reward modalities and human-like sequence planning benefit the RL approach. It is worth noting that models trained with random sequences exhibit competitive performance, particularly early in the training. This is expected, as they quickly learn to pack more objects (often smaller ones) and adapt to suitable placement strategies. In contrast, Beam-3, which is trained on human data, tends to prioritize packing larger objects first. While this approach can be more prone to failure due to suboptimal placements, Beam-3 gradually improves over time. As training progresses, it catches up and eventually surpasses other methods by learning to efficiently pack an increasing number of objects. Furthermore, while the different settings of the parameter α seem to lead to similar RL performance (with marginal improvements with CS0.6 and CS0.9), having a stability component is important for downstream robotic tasks, as it reduces the detriment of collisions between objects while placing. Qualitatively (Fig. 5), it also incentivizes a more human-like placing by keeping objects in their canonical orientations (i.e. not flipping bottles or egg cartons).

If Table V-B is considered, the benefit of the proposed approach in terms of decision latency can be observed. When compared to PackIt Heuristic [5], the shorter inference time of the proposed method is notable. Furthermore, by simplifying the RL pipeline, a significant speed-up is also obtained in training time. Still, the main bottleneck remains the simulation time, as the model itself is comparably much faster than simulating contacts in the box and ray casting to obtain the heightmap.

When examining human performance from BoxED in terms of Success Rate and Number of packed objects (Table V-B), it can be noted that in the original experiment [15] participants do not always adhere to the assumptions made by the proposed RL environment. Practically, when examining the dataset [14], some participants would place some objects leaning over the edge of the box or overcome the assumed vertical limit. This leads to somewhat worse performance than expected when the participants' performance is considered (BoxED in Table V-B). While the PackIt Heuristic SO(2) obtains comparable performance to human performance from the BoxED and to the proposed approach, it is

somewhat limited by the discretization of the environment, which leads to slightly more unreliable packing. Conversely, PackIt Heuristic SO(3)'s poor performance highlights the shortcomings in transferring heuristic assumptions to realistic scenarios. By ordering objects by volume, larger objects that are not cuboidal (egg carton, bread, bleach bottle, mustard bottle) are placed at the start of the packing somewhat non-canonically, i.e., by aligning the longest dimension with the longest side of the box. In doing so, the resulting packs rely on an unstable base, and even though the initial placements tend to be compact (Fig. 4), the later objects are unstable and fall out of the box. This is also reflected when Stability results are considered (Fig. 4). Furthermore, qualitatively, this could be deemed as poor packing since these objects could be fragile or prone to spilling. This further highlights the potential of obtaining useful information from humans implicitly, as both task performance and qualitative improvements can be achieved.

It can be observed that all the considered metrics have similar qualitative characteristics when compactness and stability are considered (Fig. 4). The stability constraint thresholded by 10° could be slightly conservative, leading to low stability metrics across the board. Furthermore, spherical objects (of which there are plenty in a household items dataset such as BoxED) tend to roll under collisions. Finally, as humans do not seem to mind these rotations while placing objects, this input could be used in designing a future system that could care about the stability of certain objects more than others.

B. Robotic System

Figure 5 shows that HERB is capable of working with data from a real depth camera in a 0-shot manner, and without any fine-tuning. The model is robust to imperfect placements and real-world physics. As expected, due to the sim-to-real gap, the execution of the model in the real world (top two rows) diverges from the same sequence run in simulation (last row). However, the prediction model remains within its learned distribution, successfully estimating poses that lead to a complete and stable packing arrangement.

While prioritizing compactness may, in theory, maximize the number of objects packed, our robotic experiments (Sec. VI) indicate that a balanced trade-off between *compactness* and *stability* yields better results ($\alpha = 0.6$ i.e. CS0.6). For example, in Fig. 5, the egg carton flips in the CS0.9 setting, which could be due to the RL agent exploiting other objects to enhance compactness, something that results in a significant hindrance to the quality of the packing.

The robotic system effectively executes sequences provided by BoxED. Human-guided sequences tend to prioritize placing stable objects first, creating a more reliable foundation for later placement of unstable or spherical objects, thereby improving overall packing stability. Importantly, by learning the continuous pose through RL the model implicitly acquired experience with challenging or potentially unstable configurations, adding a layer of reliability. Bin picking and packing in cluttered environments are very challenging problems, particularly from the dexterity point of view. By

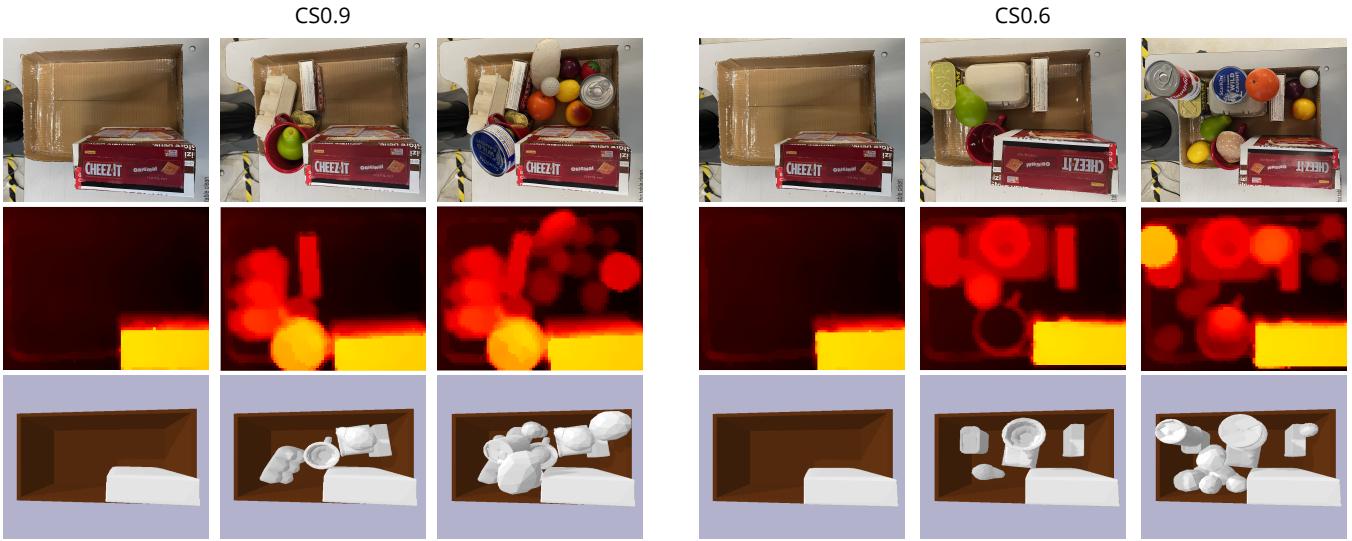


Fig. 5. Comparison of placement prediction across CS0.9 and CS0.6 parameter settings on a BoxED pack. The top two rows are the objects and the corresponding heatmap from a real sensor, the bottom is the same pack executed in simulation.

using constraints such as pre-determined grasps and top-down placements, these challenges can be mitigated but not completely alleviated. While addressing collisions and trajectory planning is out of the scope of this work, placement planning in continuous space is an important step towards more dexterous and versatile robotic agents.

VIII. CONCLUSIONS

We have shown that combining the RL approach with sequence prediction learned from humans can allow for reliable irregular object packing in 3D. Thus, the human-like sequence prediction was instrumental in reducing the complexity of the RL approach and alleviating the training time. Robotic trials were performed to validate the feasibility of the proposed methods in a practical scenario while also highlighting the importance of including a notion of stability in packing planning. Importantly, we formulated the irregular object-packing problem in a robotic setting, as a continuous control problem. Future work will focus on improving the concept of continuous bin packing and propose a more dexterous solution, possibly including grasp planning. Finally, we will also explore the use of Vision-Language Foundation Models to reason over relationships between unseen objects and generate sequences accordingly.

REFERENCES

- [1] F. Clautiaux, J. Carlier, and A. Moukrim, “A new exact method for the two-dimensional bin-packing problem with fixed orientation,” *Operations Research Letters*, vol. 35, no. 3, pp. 357–364, May 2007. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167637706000897>
- [2] G. Pantoja-Benavides, D. Giraldo, A. Montes, A. García, C. Rodríguez, C. Marín, and D. Álvarez Martínez, “Comprehensive Review of Robotized Freight Packing,” *Logistics*, vol. 8, no. 3, p. 69, Sep. 2024, number: 3 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: <https://www.mdpi.com/2305-6290/8/3/69>
- [3] A. A. S. Leao, F. M. B. Toledo, J. F. Oliveira, M. A. Carravilla, and R. Alvarez-Valdés, “Irregular packing problems: A review of mathematical models,” *European Journal of Operational Research*, vol. 282, no. 3, pp. 803–822, May 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0377221719303820>
- [4] E. G. Coffman, M. R. Garey, and D. S. Johnson, “Approximation Algorithms for Bin-Packing — An Updated Survey,” in *Algorithm Design for Computer System Design*, G. Ausiello, M. Lucertini, and P. Serafini, Eds. Vienna: Springer, 1984, pp. 49–106. [Online]. Available: https://doi.org/10.1007/978-3-7091-4338-4_3
- [5] A. Goyal and J. Deng, “PackIt: A Virtual Environment for Geometric Planning,” in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 3700–3710, iSSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v119/goyal20b.html>
- [6] S. Huang, Z. Wang, J. Zhou, and J. Lu, “Planning Irregular Object Packing via Hierarchical Reinforcement Learning,” *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 81–88, Jan. 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/9954127/>
- [7] H. Zhao, Z. Pan, Y. Yu, and K. Xu, “Learning Physically Realizable Skills for Online Packing of General 3D Shapes.” *ACM Transactions on Graphics*, vol. 42, no. 5, pp. 165:1–165:21, Jul. 2023. [Online]. Available: <https://dl.acm.org/doi/10.1145/3603544>
- [8] M. Zare, P. M. Kebria, A. Khosravi, and S. Nahavandi, “A Survey of Imitation Learning: Algorithms, Recent Developments, and Challenges,” *IEEE Transactions on Cybernetics*, vol. 54, no. 12, pp. 7173–7186, Dec. 2024, conference Name: IEEE Transactions on Cybernetics. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10602544>
- [9] “Baxter | Redefining Robotics and Manufacturing | Rethink Robotics,” Aug. 2014. [Online]. Available: <https://web.archive.org/web/20140826071530/http://www.rethinkrobotics.com/products/baxter/>
- [10] J.-H. Pan, K.-H. Hui, X. Gao, S. Zhu, Y.-H. Liu, P.-A. Heng, and C.-W. Fu, “SDF-Pack: Towards Compact Bin Packing with Signed-Distance-Field Minimization,” in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2023, pp. 10612–10619, iSSN: 2153-0866. [Online]. Available: <https://ieeexplore.ieee.org/document/10341940/?arnumber=10341940>
- [11] R. S. Sutton and A. Barto, *Reinforcement learning: an introduction*, second edition ed., ser. Adaptive computation and machine learning. Cambridge, Massachusetts London, England: The MIT Press, 2020.
- [12] H. Hu, X. Zhang, X. Yan, L. Wang, and Y. Xu, “Solving a New 3D Bin Packing Problem with Deep Reinforcement Learning Method,” Aug. 2017, publication Title: arXiv e-prints ADS Bibcode: 2017arXiv170805930H. [Online]. Available: <https://ui.adsabs.harvard.edu/abs/2017arXiv170805930H>
- [13] V. K. Chen, L. Chin, J. Choi, A. Zhang, and D. Rus, “Real- Time

- Grocery Packing by Integrating Vision, Tactile Sensing, and Soft Fingers," in *2024 IEEE 7th International Conference on Soft Robotics (RoboSoft)*, Apr. 2024, pp. 392–399, iSSN: 2769-4534. [Online]. Available: <https://ieeexplore.ieee.org/document/10521917>
- [14] A. Santos, "andrejfsantos4/BoxED," Jan. 2025, original-date: 2022-09-27T18:45:26Z. [Online]. Available: <https://github.com/andrejfsantos4/BoxED>
- [15] A. Santos, N. F. Duarte, A. Dehban, and J. Santos-Victor, "Learning the Sequence of Packing Irregular Objects from Human Demonstrations: Towards Autonomous Packing Robots," in *2024 10th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*, Sep. 2024, pp. 951–957, iSSN: 2155-1782. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10719974>
- [16] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, Jul. 2018, pp. 1861–1870, iSSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v80/haarnoja18b.html>
- [17] "Unity Real-Time Development Platform | 3D, 2D, VR & AR Engine." [Online]. Available: <https://unity.com>
- [18] "Bullet Real-Time Physics Simulation | Home of Bullet and PyBullet: physics simulation for games, visual effects, robotics and reinforcement learning." Mar. 2022. [Online]. Available: <https://pybullet.org/wordpress/>
- [19] D. Stutz and A. Geiger, "Learning 3D Shape Completion Under Weak Supervision," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1162–1181, May 2020. [Online]. Available: <https://doi.org/10.1007/s11263-018-1126-y>
- [20] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. D. Cola, T. Deleu, M. Goulão, A. Kallinteris, M. Krimmel, A. KG, R. Perez-Vicente, A. Pierré, S. Schulhoff, J. J. Tai, H. Tan, and O. G. Younis, "Gymnasium: A Standard Interface for Reinforcement Learning Environments," Nov. 2024, arXiv:2407.17032 [cs]. [Online]. Available: <http://arxiv.org/abs/2407.17032>
- [21] "Robotic Operating System." [Online]. Available: <https://wiki.ros.org/noetic>