POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Hierarchical Option-Based Reinforcement Learning for Efficient Three-Dimensional Bin Packing

Laurea Magistrale in Automation and Control Engineering - Ingegneria dell'Automazione

Author: Fausto Allegrini

Advisor: Prof. Marcello Restelli

Co-advisor: Davide Tenedini

Academic year: 2024-2025

---

1. Introduction

The 3D Bin Packing Problem (3D-BPP) is an NP-hard combinatorial optimization challenge with direct implications for operational efficiency in the logistics and supply chain industries [1]. Although the fundamental objective is to pack a set of small rectangular items into larger rectangular containers, there are numerous variations of this problem. This research specifically addresses the offline Single Knapsack Problem (SKP). In this formulation, the objective is to pack an entire given inventory of items (called an "order") into a single bin, minimizing the volumetric occupancy. Operating in an offline paradigm, the decision-making agent has full prior knowledge of the order before the packing sequence begins. To ensure physical realizability in real-world industrial scenarios, the proposed packing framework enforces a strict set of operational constraints. These include geometric containment within the bin boundaries, strict non-overlapping of items, orthogonal alignment with the container axes, and restricted orientation (allowing rotation exclusively around the vertical z-axis). The system also enforces rigorous static stability to prevent structural collapse under gravity. Finally, to satisfy packing complexity constraints, the generated arrangements must be kinematically executable by a robotic manipulator without causing collisions with previously packed items.

To overcome the computational limitations of traditional heuristic and meta-heuristic search algorithms, the state-of-the-art has increasingly transitioned toward Deep Reinforcement Learning (DRL).

As a foundational step, this thesis provides a systematic taxonomy of existing DRL-based solutions for the 3D-BPP, identifying a significant gap: existing learning-based methods frequently abstract away the physical occlusions and reachability limits of real robotic manipulators and often do not assess the structural stability of the loaded items.

To address these limitations, this research proposes a novel architecture that integrates a Hierarchical Reinforcement Learning (HRL) framework with deterministic geometric heuristics, aiming to generate robust and strictly feasible packing sequences.

2. Survey of DRL-Based 3D-BPP Solvers

The survey was conducted on 26 recent publications applying DRL to the 3D Bin Packing Problem. The examined literature was initially partitioned by problem formulation, identifying 9 papers addressing the offline 3D-BPP and 17 focusing on the online variant. The proposed taxonomy categorized these methodologies based on their core reinforcement learning components: state and action space representations, reward shaping strategies, algorithmic frameworks, and the chosen neural network architectures for function approximation.

Furthermore, the analysis evaluated how these works decompose the combinatorial action space through the integration of deterministic heuristics, and which criteria are used to assess structural stability. This review identified a distinct gap in the current state-of-the-art, particularly among offline solvers. Existing methodologies systematically omit practical feasibility requirements, failing to enforce rigorous static stability and ignoring the physical occlusions that limit the reachability of real-world robotic manipulators. Additionally, while the integration of heuristics proves effective in reducing problem complexity, current approaches often impose rigid spatial constraints that overly restrict the learning agent's exploration. The architecture proposed in this thesis was specifically designed to address these structural and kinematic limitations.

3. Problem Statement

The offline 3D-BPP is inherently a sequential decision-making process. At each time step t, the placement of a new item depends solely on the current spatial topology of the container and the remaining inventory. Consequently, the problem is formally modeled as a Constrained Markov Decision Process (CMDP).

3.1. Environment State

The global state of the environment is defined as $S_t = \langle s_t^u, \mathcal{C}_t \rangle$, where s_t^u represents the unpacked inventory and $\mathcal{C}_t$ denotes the active container configuration.

Unpacked Items State. To efficiently manage varying quantities of items and natively embed rotational degrees of freedom, the unpacked inventory s_t^u is structured as a set of virtual oriented clusters. Each physical item type is split into two distinct clusters corresponding to its valid orthogonal horizontal orientations. These paired clusters share a quantity counter, ensuring that the selection of an item in either orientation correctly updates the global physical inventory.

$$s_t^u = \{\nu_{1,0}, \cdots, \nu_{N_k,1}, \},$$
$$\nu_{k,o} = (w_k(o), l_k(o), h_k, q_k, o)$$

Container State ($\mathcal{C}_t$). The container state $\mathcal{C}_t$ is decomposed into an observable spatial feature map s_t^c and a set of heuristic action masks $\mathcal{M}_t^c$ in one-to-one relation with the items of s_t^u. Following the methodology proposed in [4], the 3D container volume is projected into a 5-channel 2D spatial tensor $s_t^c \in \mathbb{Z}^{L \times W \times 5}$. The primary channel encodes the current maximum stacked height (the heightmap). The remaining four channels encode specific plane features, explicitly measuring the contiguous distance from any grid coordinate to the nearest topographic discontinuity or edge along the four cardinal directions (see [Fig.1]).

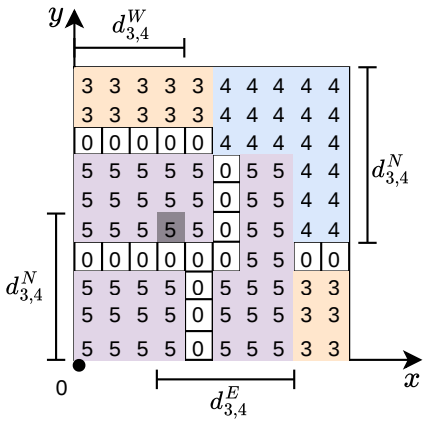

Figure 1: Example of a 10×10 feature map

The heuristic action masks $\mathcal{M}_t^c$ translate geometric and stability conditions into boundaries for the learning agent. For each oriented cluster, a corresponding binary spatial mask evaluates the topography confining the action space to heuristically determined placements.

Geometric Approximation: The Pseudo-Planar Support. In real-world scenarios, small height differentials (steps) and narrow depressions (gaps smaller than the minimum available item footprint) can still provide secure support. Introducing the approximation of a pseudo-planar support, the raw heightmap is geometrically filtered to bridge these negligible gaps. This yields a pseudo-planar heightmap $\mathcal{H}^m_{plane}$ that expands the set of structurally viable placement coordinates, increasing the achievable volumetric density without violating stability constraints.

3.2. Feasibility and Rigorous Stability Validation

To enforce the static stability constraint, the environment implements a rigorous validation based on the *Load Bearable Convex Polygon* (LBCP) framework [2]. An item is considered stable if its Center of Mass (CoM) region lies entirely within its designated support polygon $P^{\triangle}_{\text{new}}$.
The support polygon is computed leveraging the pseudo-planar heightmap. The new support polygon is defined as the Convex Hull (CH) of the intersection between the item's pseudo-planar contact points ($\text{PS}_{\text{contact}}$) and the structurally feasible regions ($\text{PS}_{\text{feasible}}$) extracted from the global feasibility map $\mathcal{F}^m_t$:

$$P^{\triangle}_{\text{new}} \leftarrow \text{CH}(\text{PS}_{\text{contact}} \cap \text{PS}_{\text{feasible}}) \tag{1}$$

The placement is validated if and only if $C_{\text{new}} \subseteq P^{\triangle}_{\text{new}}$.
Upon validation, the global feasibility map is updated through a Structural Stability Update (SSU) to mask out non-load-bearing regions (cantilevers).

3.3. Action Space Representation

The decision-making process is decomposed into selecting an oriented item from s^u_t and determining its planar Loading Position (x, y). Since the vertical placement (z-coordinate) is deterministically established by the current heightmap (bottom-up packing), the theoretical 3D action space is effectively reduced to a 2D grid:

$$\mathcal{A}_{\text{eff}} = 2N_k \times L \times W. \tag{2}$$

Heuristic constraints are enforced directly on this space by $\mathcal{M}^c_t$, deterministically masking out any non-feasible spatial coordinate (x, y).

3.4. Reward Function

Learning is guided by a step-wise reward formulated with two components:

- **Gap Reduction (r^{gap}_t):** Rewards the step-by-step reduction of the bounding empty space g_t, mathematically driving the policy to fill existing voids before increasing the overall pile height $\tilde{H}_t$:

$$\begin{aligned} r^{\text{gap}}_t &= g_{t-1} - g_t, \\ g_t &= \left(LW\tilde{H}_t\right) - \sum_{i \in \mathcal{P}_t} l_i w_i h_i. \end{aligned} \tag{3}$$

- **Dead Space Penalty (r^{dead}_t):** Penalizes the creation of permanently inaccessible voids by analyzing the topological consequences of the placement. If a placement irreversibly blocks lanes or covers depressions smaller than the minimum item footprint ($w_{min} \times w_{min}$), the resulting empty volume is flagged as wasted space.

The final scalar reward R_t aggregates these components via weighting hyperparameters (w_g, w_d), applying a severe fixed penalty for statically unstable actions:

$$R_t = \begin{cases} w_g r^{\text{gap}}_t + w_d r^{\text{dead}}_t & \text{if } a_t \text{ stable;} \\ -\dfrac{L \times W}{50} & \text{if } a_t \text{ unstable.} \end{cases} \tag{4}$$

4. Methodology

To solve the 3D-BPP, this research decomposes the packing process into a layer-building strategy using Hierarchical Reinforcement Learning (HRL). By extending the CMDP into the Options framework [5], the atomic placement of items is abstracted into temporally extended macro-actions responsible for constructing entire horizontal layers. Furthermore, targeted heuristics are integrated to streamline the state-action space and enforce stability.

4.1. Hierarchical Framework and The Options Logic

The decision-making process is structured HRL architecture based on the Options framework. This approach abstracts the atomic placement of single items into temporally extended macro-actions.
The option space $\Omega(S_t)$ available to the agent at decision time t is defined as:

$$\Omega(S_t) = \{\omega(h) \mid h \in \{0\} \cup H_{cand}(S_t)\} \tag{5}$$

where $h = 0$ triggers a primitive, single-step placement ("Free Mode"), and $h > 0$ triggers a temporally extended layer-building macro-action ("Layer Mode") parameterized by the target layer height h.

High-Level Policy (Heuristic Manager). The Manager is designed as a fast, heuristic-driven module. It first evaluates the pseudo-planarity of the current placement surface. If the surface is pseudo-planar, the Manager calculates the potential horizontal coverage area A_h for each candidate height subset s_h^u available in the inventory:

$$A_h = \sum_{\nu_k \in s_h^u} (l_k \times w_k \times q_k) \tag{6}$$

It then greedily selects the optimal height parameter $h^* = \arg\max_h(A_h)$ to maximize volumetric density and initiates the corresponding option $\omega(h^*)$. If the surface is not pseudo-planar, it defaults to the primitive action $\omega(0)$.

Low-Level Policy (Shared Worker). The Worker utilizes a single shared policy $\pi_w(a_t \mid S_t, \omega)$ across all options. Behavioral differentiation between options is achieved through *Conditional Action Masking*. During a layer option $\omega(h)$, the dynamically generated mask $\mathcal{M}_h^c$ restricts item selection strictly to the subset s_h^u and locks the valid placement coordinates to the exact elevation of the targeted pseudo-planar surface z_{curr}.

4.2. Heuristic Construction of the Action Mask

To constrain the spatial action space without over-restricting the learning agent, this research introduces **Corner-Anchored Regions (CAR)**. CAR extends the corner points (CPs) heuristic into continuous planar regions by anchoring a fictitious minimum-sized box footprint ($w_{min} \times w_{min}$) at each CP. A global preliminary mask $\mathcal{M}_{\text{CAR}}$ is constructed as the boolean union of these regions, preventing the generation of unfeasible coordinates while maintaining spatial versatility for the policy.

To guarantee physical realizability, this preliminary mask is filtered through additional geometric and structural constraints. First, a containment mask $\mathcal{M}_{\text{cont}}^{(k,o)}$ deterministically discards any coordinate that would cause the selected item to protrude beyond the container's boundaries. Second, an edge-alignment heuristic mask $\mathcal{M}_{\text{edge}}^{(k,o)}$ is applied to prevent the creation of unfilled gaps near the container walls, forcing items placed near the boundaries to align perfectly with the edges.

Furthermore, a vectorized fast stability check is introduced to generate a stability mask $\mathcal{M}_{\text{stab}}^{(k,o)}$ using the plane features and the global feasibility map $\mathcal{F}_t^m$. It heuristically guarantees static stability by verifying that the item's Center of Mass (CoM) is adequately supported by the pseudo-planar surface along both spatial axes and rests on a structurally viable region.

The final valid placement mask $\mathcal{M}_t^{c,(k,o)}$ is computed as the logical intersection of the masks for each oriented physical item, mathematically:

$$\mathcal{M}_t^{c,(k,o)} = \mathcal{M}_{\text{CAR}} \cap \mathcal{M}_{\text{edge}}^{(k,o)} \cap \mathcal{M}_{\text{stab}}^{(k,o)} \cap \mathcal{M}_{\text{cont}}^{(k,o)} \tag{7}$$

$\mathcal{M}_{\text{t}}^c$ is the set of all the $\mathcal{M}_t^{c,(k,o)}$ masks. By applying this combined boolean filter, the architecture zeroes out all structurally unsound or geometrically invalid actions, and heuristically limits the action space.

4.3. Worker Policy Architecture

The unified Worker policy π_w is structurally decomposed into an item selection sub-policy $\pi_{sel}(idx \mid S_t, w)$ and a spatial placement sub-policy $\pi_{pla}(x, y \mid idx, S_t, w)$. Both components are approximated using specialized Deep Neural Networks designed for spatial reasoning and representation learning.

The state representations are initially processed by dedicated encoders. The discrete inventory state is embedded and passed through a Transformer Encoder. To provide immediate spatial awareness, these embeddings are augmented with dynamic feasibility features derived directly from the option-specific action mask. Simultaneously, the container's multi-channel spatial feature map is processed by a hybrid CNN-Transformer module. The CNN extracts local topographic features and downscales the spatial resolution to reduce computational overhead. The flattened sequence is then processed by a Transformer Encoder, which models long-range dependencies and aggregates the global container context into a single classification (`[CLS]`) token.

The selection sub-policy identifies the optimal item and its horizontal orientation using a Pointer Network. This module performs masked cross-attention between the spatial `[CLS]` token and the encoded inventory, outputting a discrete probability distribution over the available items. Conditioned on this selection, the placement sub-policy computes the exact spatial coordinate for the item. It utilizes a Transformer Decoder to cross-attend the explicitly selected item's embedding with the downscaled spatial feature map. To map these decisions back to the native container resolution, the resulting coarse logits map is upsampled via bilinear interpolation. Finally, a masked Softmax layer applies the action mask $\mathcal{M}_h^{c,(k,o)}$, deterministically zeroing out unfeasible coordinates to yield a strictly valid 2D spatial probability distribution.

4.4. Critic Function Approximator

To stabilize the optimization of the Worker policy within the Proximal Policy Optimization (PPO) framework, a Critic network is employed to estimate the state value function $V(S_t)$. To synthesize a comprehensive understanding of the environment, the Critic aggregates three distinct information streams: the global spatial `[CLS]` token from the Heightmap Encoder, the aggregated latent representation from the Item Encoder, and a highly condensed six-dimensional vector of macroscopic state heuristics.
These explicit heuristics include the logarithmic count and aggregate volume of the remaining inventory, the mathematical mean, maximum, and standard deviation of the container's heightmap, and a boolean indicator signaling the presence of a pseudo-planar surface. The concatenated state vector is processed through a Multi-Layer Perceptron. The network outputs a single continuous scalar representing the expected return, establishing a robust baseline for variance reduction during the policy gradient updates.

5. Results

The proposed HRL model is evaluated on a test dataset comprising 1000 distinct packing orders, statistically generated from real-world industrial logistics data, to be loaded onto an industrial Euro-pallet bin. To manage computational complexity during training, the spatial dimensions are scaled down, resulting in a discrete container grid of 24×16 units with a maximum height of 32 units.
The HRL agent was trained exclusively in a restricted *single-layer* setup, where episodes terminate immediately upon completion of the first horizontal layer. This experimental design is deliberately chosen to isolate the layer-building Option and drastically reduce the computational burden of training over the full 3D combinatorial space. The resulting policy is evaluated in both the single-layer setup and the full *unrestricted bin* environment. This strategy tests the architecture's zero-shot generalization capabilities. The proposed model is benchmarked against two established baselines: a GA employed as solver in a real logistics company and a standard basic heuristic based on Corner Points.

5.1. Comparative Analysis: Single-Layer Setup

For this evaluation, two KPIs are considered: the percentage of orders in which a layer was successfully generated, and for these instances, the Area Utilization of the loaded items in the layer.
The HDRL framework successfully completed a pseudo-planar layer in 79.22% of the instances, achieving a mean area utilization of 83.67%. While the GA established the empirical upper bound for spatial efficiency (80.76% success rate, 88.17% area utilization), the execution of the HDRL model takes an average of an average of 1.10, which is 9 times faster than the GA (which requires 9.10 seconds). The standard heuristic baseline proved structurally unsuited for targeted layer-building, fulfilling the pseudo-planar condition in only 17.81% of the evaluated orders.

Metric	GA	HDRL	O3DBP
Success Rate	**80.76%**	79.22%	17.81%
Area Util.	**88.1% ± 6.3%**	83.6% ± 5.9%	63.3% ± 5.3%
Exec. Time	9.10s ± 16.7s	**1.10s ± 0.1s**	3.14s ± 1.1s

Table 1: Performance comparison of the *layer_ option* across 1000 evaluated orders.

5.2. Generalization Analysis: Unrestricted Bin Environment

The generated packing configurations were evaluated using standard KPIs, including volume utilization, physical stability (validated through

rigid-body physics simulations) and an overall algorithmic benchmark score proposed by [3].
The model was tested under two operational settings: **HDRL (Rigorous)**, which enforces the convex hull support validation, and **HDRL (Fast-Check)**, which does not employ the convex hull verification.
As detailed in Table 2, the HDRL (Rigorous) configuration maintained physical stability in 99.9% of the test cases. Notably, the HDRL (Fast-Check) achieved a highly reliable 95.3% stability rate, confirming the effectiveness of the heuristic masking strategy. While the Fast-Check setting yielded slightly higher volume utilization (66.4% vs. 64.2%) due to increased spatial freedom, the Rigorous setting achieved a higher overall score by strictly eliminating algorithmic penalties for unstable placements.

Model Configuration	Volume Util.	Final Score	Stable Orders	Exec. Time
GA Baseline	**75.7% ± 7.1%**	**0.728**	**100.0%**	717.4s ± 385.7s
HDRL (Fast Check)	66.4% ± 8.4%	0.576	95.3%	**2.54s ± 0.5s**
HDRL (Rigorous)	64.2% ± 8.5%	0.598	99.9%	**2.54s ± 0.5s**
Heuristic O3DBP	54.5% ± 3.8%	0.002	0.3%	113.9s ± 30.0s

Table 2: Performance comparison on the full 3D-BPP environment.

Compared to the baselines, the standard CP heuristic completely fails to produce stable configurations (0.3% success rate). The Genetic Algorithm establishes the empirical upper bound for volume utilization (75.73%); however, it requires an average of nearly 12 minutes (717.48 seconds) per order. In stark contrast, the proposed HDRL architecture successfully generates a complete, multi-layer, and physically stable packing plan in just 2.54 seconds. This represents a computational acceleration by a factor of nearly 300 compared to the GA baseline.

6. Conclusion

This research successfully bridged the gap between theoretical DRL and the strict physical constraints of industrial robotic packing. A comprehensive taxonomy of DRL literature within the 3D-BPP domain was developed, systematically identifying methodological gaps regarding structural stability and robotic reachability. To overcome these limitations, this thesis introduced a novel HRL architecture. This framework effectively reduces combinatorial complexity through a practical "pseudo-planar" layer approximation and strictly enforces kinematic realizability by integrating deterministic heuristics with a Corner-Anchored Regions (CAR) action masking mechanism.
The experimental results provide encouraging evidence of the applicability of the proposed methodology in real-world industrial scenarios.
It is demonstrated that the approach of combining DRL with a layer-building pattern is valid, as even on a dataset of orders with a limited number of items, the model constructs a layer in 79.22% of instances.
Moreover, the hierarchical decomposition of the learning policy exhibits strong generalization capabilities; specifically, the model trained exclusively for single-layer construction successfully operates within the full 3D environment, outperforming the basic Corner Point-based heuristic.
A significant finding concerns the efficacy of the action masking mechanism based on heuristic stability. The validity of this approach is confirmed by the attainment of stable configurations in 95.3% of the tested orders (when not supported by the rigorous stability check), while simultaneously preserving sufficient degrees of freedom to enable learning an efficient policy.
This balance, coupled with the superior performance of the HDRL architecture compared to the heuristic algorithm and the significantly reduced computational times, establishes a robust solution for automated logistics.

References

[1] Sara Ali, António Galrão Ramos, Maria Antónia Carravilla, and José Fernando Oliveira. On-line three-dimensional packing problems: A review of off-line and on-line solution approaches. *Computers & Industrial Engineering*, 168:108122, 2022.

[2] Ziyan Gao, Lijun Wang, Yuntao Kong, and Nak Young Chong. Online 3d bin packing with fast stability validation and stable rearrangement planning. *ArXiv*, abs/2507.09123, 2025.

[3] Florian Kagerer, Maximilian Beinhofer, Stefan Stricker, and Andreas Nüchter. Bed-bpp: Benchmarking dataset for robotic bin packing problems. *The International Journal of Robotics Research*, 42(11):1007–1014, 2023.

[4] Quanqing Que, Fang Yang, and Defu Zhang.

Solving 3d packing problem using transformer network and reinforcement learning. *Expert Systems with Applications*, 214:119153, 2023.

[5] Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211, 1999.