

Learning to Optimize Package Picking for Large-Scale, Real-World Robot Induction

Shuai Li¹, Azarakhsh Keipour¹, Sicong Zhao¹, Srinath Rajagopalan¹,
Charles Swan¹, and Kostas Bekris^{1,2}

¹ Amazon Robotics, Seattle WA 98109, USA,
{amzshua, keipourv, sczhao, rajsrina, cswan, bekris}@amazon.com,
² Rutgers University, Piscataway, New Jersey 08854, USA

Abstract. Warehouse automation plays a pivotal role in enhancing operational efficiency, minimizing costs, and improving resilience to workforce variability. While prior research has demonstrated the potential of machine learning (ML) models to increase picking success rates in large-scale robotic fleets by prioritizing high-probability picks and packages, these efforts primarily focused on predicting success probabilities for picks sampled using heuristic methods. Limited attention has been given, however, to leveraging data-driven approaches to directly optimize sampled picks for better performance at scale. In this study, we propose an ML-based framework that predicts transform adjustments as well as improving the selection of suction cups for multi-suction end effectors for sampled picks to enhance their success probabilities. The framework was integrated and evaluated in test workcells that resemble the operations of Amazon Robotics' Robot Induction (Robin) fleet, which is used for package manipulation. Evaluated on over 2 million picks, the proposed method achieves a 20% reduction in pick failure rates compared to a heuristic-based pick sampling baseline, demonstrating its effectiveness in large-scale warehouse automation scenarios.

Keywords: Robotic Manipulation at Scale, Pick Optimization, Learned Picking, Warehouse Robots

1 Motivation, Problem Statement and Related Work

Robot induction systems, such as the one shown in Fig. 1, which automate singulation and sorting in logistics, face multiple challenges. They include high object diversity, deformable material, complex pile configurations, and significant occlusions. Despite significant progress in computer vision and machine learning, identifying effective picks remains a critical bottleneck for effective, real-world industrial operations [1]. In particular, improve-



Fig. 1: An example of a Robot induction (Robin) workcell. Robin performs automated package singulation by picking packages from unstructured piles on conveyors and transferring them to mobile robots.

ments in pick success rates can translate to significant operational gains in large-scale industrial settings.

Prior Work in Robot Picking Traditional methods for robot grasping [13] involve geometric reasoning and planning typically for multi-fingered hands. They often calculate the poses and forces for robotic contacts to satisfy mechanical constraints. The requirement, however, for accurate geometric and physical object models often limits their applicability in unstructured setups involving open-sets of objects [2].

This has motivated the development of learning-based picking approaches [7]. Such data-driven methods brought the promise of higher effectiveness in challenging, cluttered setups with unknown objects. Morrison et al. [15] estimated grasp quality from depth images, while Jiang et al. [14] explored geometric, feature-based approaches. Most ML methods are focused on parallel, pinch grippers [5, 6, 17] or are dealing with simple setups, such as single-object picking. The Amazon Picking Challenge [4], however, demonstrated that suction-based grippers can be very effective in real-world picking setups by simplifying pick reasoning. Suction-based grippers have been deployed in real-world production environments and allow fast and robust picking of items that can potentially be heavy.

Some of the data-driven solutions for robot picking have been extended to address suction-based grippers, such as the Dex-Net solutions, which have addressed open-ended grasp quality prediction from point clouds [12, 3]. In Dex-Net, a pick candidate is evaluated using an expert-designed evaluation system. These promising approaches, however, may not fully address the complexity and tight operational constraints of industrial settings. Researchers have explored interactive strategies for scenarios where picks are not immediately available [11], but these approaches tend to be too time-consuming for practical applications.

The current work assumes access to a model that has learned pick success prediction given success labels of past production data [8, 9]. This prior work by Li et al. has explored shallow, explainable models trained on large-scale, real-world data. More recent work in learning pick success models for multi-suction picks [16] has shown that with proper pretraining and finetuning, a deep model with a multimodal visual encoder can in fact outperform the shallow model as well as alternative deep architectures in terms of picking items from an open-set.

Contribution This work presents a novel approach that focuses on optimizing candidate picks given access to a model of pick success probability. The input picks can be generated either via existing heuristic or learned solutions. The approach refines the parameters of the input picks by using a machine learning model that is trained on small-scale data gathered from physical robots and utilizing the learned model of pick success. The proposed optimization model outputs higher-quality picks that improve success rates. We demonstrate effectiveness via comprehensive evaluation over two million robotic inductions performed on test workcells that resemble Amazon warehouse operations for robot induction.

Overall, the key contributions include: (1) **Pick Optimization Framework:** A novel method for adjusting picks to enhance their success probability given access to few physical data and a prior model for evaluating pick success; (2) **Demonstrable Improvement:** Enhanced performance over ranking-based approaches based on the same

underlying model for evaluating pick success that rank multiple samples, validated through extensive A/B testing involving 2 million picks.

2 Technical Approach

Consider the picking task illustrated in Fig. 1. The task is initiated when a picking scene s_t of cluttered packages of different types arrives at a reachable area via a conveyor belt. The conveyor belt and the scene remain static throughout the picking process. The picking is performed by an induction manipulation robot consisting of a multiple-DoF arm with a end-of-arm tool (EoAT) that can activate multiple suction cups.

Each pick action a is defined as a set of variables determining: a 3-D point in space (i.e., the desired pick point where the EoAT makes contact with an item’s surface), the desired 3-D orientation of the EoAT at the pick point, and a set of desired active suction cups on the EoAT.

A feature vector $\phi(s_t, a) \in \mathbb{R}^d$ encodes both pick action a parameters and scene s_t information, such as the geometric relationship between the end-of-the-arm tool (EoAT) and nearby surfaces. A function $F : \mathbb{R}^d \rightarrow [0, 1]$ maps these features $\phi(s_t, a)$ to pick success probability.

Access to Learned Model of Pick Success: Previous research demonstrated machine learning models that approximate the pick success probability function F using historical data from executed picks in industrial settings [8]. This work assumes access to such learned models of F .

Optimization Objective: For an initial candidate pick a , its local neighborhood $\mathcal{N}(a)$, and the scene s_t , our approach incrementally refines the pick parameters to maximize the predicted success probability, which can be expressed as finding a^* as follows:

$$a^* = \arg \max_{a' \in \mathcal{N}(a)} F(\phi(s_t, a'))$$

Features: The features $\phi(s_t, a)$ can be shared between the model of pick success $F(\phi(s_t, a'))$ and the proposed optimization model that refines pick parameters. We compute a set of features for each induct using scene metadata as well as RGB, and depth images. Specifically, the camera data is processed to generate package segments and tag each segment with an associated package type label. Additional statistics are computed for each segment using depth information (e.g., surface normals and the quality of plane fitting). The following features have been shown to be good predictors of pick quality:

- Package height: This feature correlates with the package’s momentum and can impact the shear force at the suction cups and the pick’s stability.

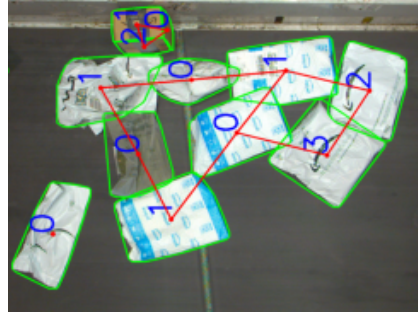


Fig. 2: Example of an adjacency graph for a cluster of items.

- Quality of plane fitting: A plane is fit on each segment; a better plane fit correlates with a better seal between the suction cups and the package.
- Number of activated suction cups: More active suction cups means a more stable pick, reducing the failure probability.
- Alignment quality between the suction cups and the package surface: This feature is computed as the offsets between the package surface normal vector and the normal vector of the suction cups. A better alignment indicates a better seal between the suction cups and the package surface.

In addition to the above and other segment-specific features, we compute features that describe each segment’s relationship with its surroundings, including the number of nearby segments, the *adjacency graph* and the *local height map* features. To compute the adjacency graph features, we construct a graph that captures the topological order of the package segments. This graph captures each detected segment’s relative height with respect to its adjacent neighbor segments. Figure 2 shows an example where the numbers represent the relative position ranking of the segment among its neighbors. To compute the local height map, we select point cloud within d meters vicinity of the target package, and compute the height map with the coordinates of the points along the world Z axis. The local height map provides detailed spatial and geometry context of the target package as well as its surrounding environment. Figure 3 shows the cropped RGB image of an package and the corresponding local height map.

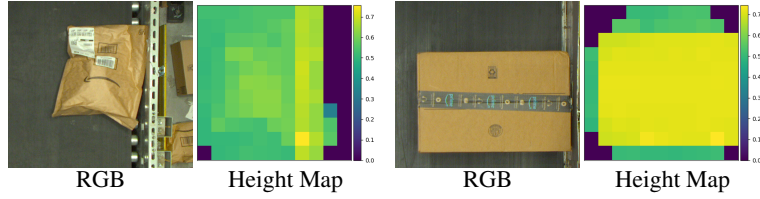


Fig. 3: Examples of local height maps of packages.

Training Data Generation: The training data is generated via a 2 step process:

- First, we apply controlled noise to an executed pick action a_i to obtain a perturbed action a_{i+1} ;
- Then, we observe the success probability using model F both before and after applying the noise:

$$p_i = F(\phi(s_t, a_i)), \quad p_{i+1} = F(\phi(s_t, a_{i+1})).$$

To generate the perturbed action a_{i+1} , we sample Gaussian noise ε_a with zero mean and predetermined variances (σ_{pos} for positions and σ_{rot} for rotation).

For each pair consisting of a perturbed action a_{i+1} and its original pick action a_i , we define a noise term δ_i that represents the desired gradient for optimizing pick success probability.

$$\delta_i = \begin{cases} a_i - a_{i+1}, & \text{if } p_i > p_{i+1} \\ a_{i+1} - a_i, & \text{otherwise} \end{cases} \quad (1)$$

Each noise term is paired with a corresponding feature vector ϕ_i :

$$\phi_i = \begin{cases} \phi(s_t, a_{i+1}) & \text{if } p_i > p_{i+1} \\ \phi(s_t, a_i) & \text{otherwise} \end{cases} \quad (2)$$

The pairs (ϕ_i, δ_i) constitute the training dataset derived from 1,000 robot picks sampled from real-world picking experiments in testing workcells. For each real-world executed pick a_i , we also generate N perturbed actions by adding Gaussian noise. This process yields the feature and target pairs described above. The resulting dataset contains 27,977 data points split into 22,381 training data points and 5,596 testing data points.

Figure 4 illustrates the described data generation approach.

Model Architecture and Training This work trains regression models G that predict action noise/gradient terms δ based on features $\phi(s_t, a)$. We employ an autoregressive approach that predicts each pick action dimension separately, while considering information from other pick action dimensions. This results in 3 distinct models corresponding to the 3 dimensions of the pick action: X and Y coordinates of the EoAT, and the rotation angle r about the vector normal to the package surface. Note that the normal vector and Z -coordinate of the pick are computed based on the surface below the EoAT at the given X and Y coordinates.

The models are structured as follows: the X -axis translation model G_x uses the base features ϕ_i ; the Y -axis translation model G_y incorporates both ϕ_i and the predicted X value; and the angular rotation model G_r utilizes ϕ_i along with predicted X -axis and Y -axis values. This sequential prediction strategy ensures that each subsequent prediction

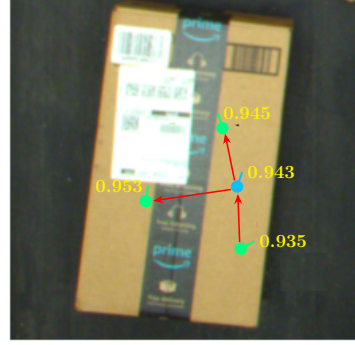


Fig. 4: An illustration of the data generation approach. An initial action a_i (blue) and the N generated perturbed actions (green) are shown along with their predicted success probability scores. The vectors correspond to the noise term δ_i that the proposed model is learning.

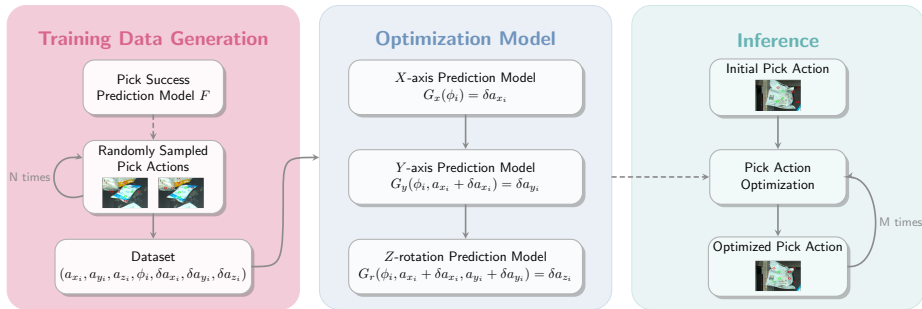


Fig. 5: The overview of the system.

can leverage information from previously predicted parameters, potentially capturing important spatial relationships in the picking task.

Figure 5 illustrates the design overview of the described system.

3 Experiments

We evaluated the proposed pick optimization method via using test workcells that resemble real-world logistics operations in fulfillment centers. This section details the experimental setup, presents quantitative results, and analyzes our findings.

Robotic System Architecture The test workcells resemble logistics operations, such as Amazon’s Robin robotic system [1], which has been developed for package manipulation (Fig. 6). The test workcells make use of a FANUC M-20iD/35 industrial robotic arm with six degrees of freedom, providing maximum operational flexibility, a 35 *kg* maximum payload capacity, and a 1831 *mm* operational reach. The end-of-arm tool (EoAT) features a vacuum-based system with 8 independently controllable suction cups arranged in an “X”-pattern configuration, covering a 25 *cm* × 25 *cm* effective grasp area.

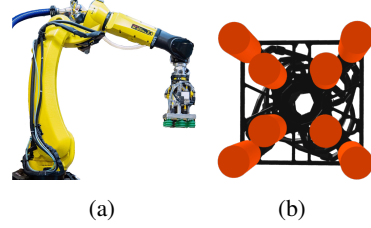


Fig. 6: Amazon’s Robin configuration used in our experiments. (a) Arm. (b) Bottom view of the EoAT.

Model Architecture Our study evaluates and compares two distinct model architectures: a gradient boosting algorithm and a 2-layer Multi-Layer Perceptron (MLP). The MLP architecture consists of two hidden layers with 35 and 2 neurons respectively, and utilizes default parameters, including the Adam optimizer with a learning rate of 0.001. We assess both models using the Root Mean Square Error (RMSE) metric on the testing dataset, calculating the RMSE for each of the three pick parameter dimensions. The comparative performance results of both models are presented in Table. 1.

Pick Parameter	Grad. Boosting	2-Layer MLP
p_x (meters)	0.0248	0.0277
p_y (meters)	0.0254	0.0319
r_z (radians)	0.2971	0.3293

Table 1: RMSE Comparison Across Models

Quantitative Results To evaluate the effectiveness of our learned pick optimization approach, we conducted comprehensive A/B testing experiments comparing two distinct groups. *Control group (C)*: Baseline performance using our nominal heuristic-based pick planning algorithm with a learned pick ranker (similar to Li et al. [8]). *Treatment group (T)*: Enhanced performance using our optimized models and algorithms.

Group	Missed Picks	Missed Picks Mean Rate	Missed Picks 95% CI
C	22,310	2.23%	[2.23%, 2.23%]
T	18,015	1.80%	[1.79%, 1.81%]

Table 2: Comparison across Treatment Groups (1M inducts per treatment group)

One million picks from testing workcells were used for evaluation purposes for each of the treatment groups, i.e., 1 million picks for group C and 1 million picks for group T. The primary evaluation metric was missed pick rate, which measures the frequency of failed package transfers from the conveyor belt to the designated destination region. Table 2 provides the corresponding statistics indicating a significant, statistically significant reduction of missed pick rate by the proposed pick optimization process.

In addition to the missed pick rate, we evaluated two additional metrics 1) infeasible grasp rate, when none of the generated grasps are feasible due to motion feasibility or collision, and 2) multi-pick rate, when the robot picks multiple items simultaneously. The results of these two metrics are shown in Table 3.

Group	Infeasible Grasp Mean Rate	Infeasible Grasp 95% CI	Multi-pick Mean Rate	Multi-pick 95% CI
C	4.49%	[4.49%, 4.50%]	0.84%	[0.84%, 0.85%]
T	4.48%	[4.46%, 4.49%]	0.89%	[0.89%, 0.90%]

Table 3: Comparison across Treatment Groups (1M inducts per treatment group)

4 Insights

Experimental Scale: The evaluation involved a comprehensive A/B experiment by utilizing 2 million physical picks from testing workcells that resemble typical setups that are operational across the logistics industry, such as Amazon’s Robin system. To the best of the authors’ knowledge, the performed experiment represents one of the *largest-scale evaluations ever reported on physical robotic manipulation hardware*.

Qualitative Results: Figure 7 provides qualitative examples from the physical evaluation of the proposed machine learning approach for optimizing robotic pick parameters in manipulation tasks. The method draws inspiration from flow matching techniques [10], where machine learning models learn to map initial pick parameters to their optimal values through a learned vector field. The proposed strategy learns such a vector field via supervision given a previously trained Pick Success Prediction (PSP) model F . The proposed framework could be adapted to accommodate alternative objectives, such as incorporating human preferences, through modifications to the training data generation process, such as leveraging human annotations and scores on the grasps.

The examples above highlight that training a vector field in this manner allows the proposed method to iteratively refine the initial sampled pick actions (left images) to progressively improve their predicted success probabilities (after 2 and 3 iterations). On the top row, the grasps always activate 3 suction cups but the position of the suction cups is incrementally adapted by the model so that eventually the cups are pressed against surfaces of the package with improved attachment - and away from the label, which may be ripped off by the suction cup during transfer. On the bottom row, the initial sampled grasp had only 2 suction cups activated but the model is able to optimize the pick to allow for 3 active suction cups, which typically means a more robust pick.

Statistical Significance and Impact on Operations: The evaluation focuses on analyzing missed pick events, which occur when packages fail to be successfully placed on mobile robots, as detected by an integrated perception system. The A/B experiments

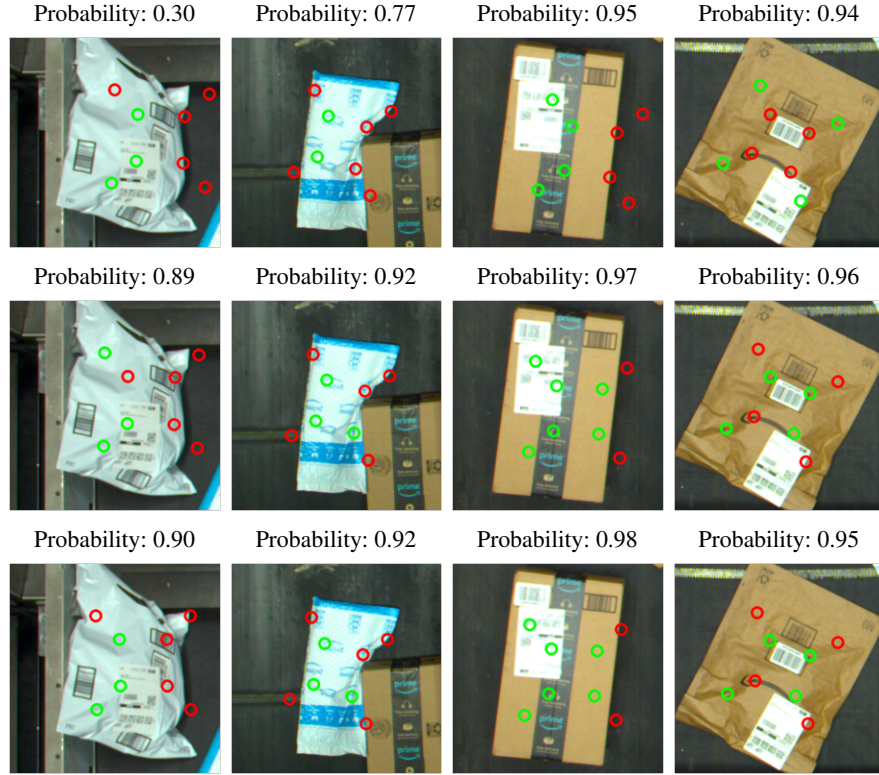


Fig. 7: Qualitative examples of pick optimization. The circles show the location of suction cups projected over the scene. Red is for retracted, deactivated suction cups. Green corresponds to extended, activated cups. The probabilities on top of each image correspond to the reported probabilities estimated of the Pick Success Prediction (PSP) model F . (first row) Initially sampled pick action. (second row) Optimized action after 2 iterations of the proposed approach. (third row) Optimized action after 3 iterations of the proposed approach.

revealed statistically significant reduction in missed pick rates compared to alternatives, underscoring the potential of pick optimization via ML in improving real-world robotic manipulation performance over time. When a missed pick occurs, one of two scenarios unfolds: (i) either the package remains or falls onto the conveyor, requiring a re-pick attempt and reducing the workcell’s throughput and operational efficiency, or (ii) it results in an “amnesty” case when it is dropped outside the conveyor belt. Amnesty cases incur even greater costs due to the necessity of human intervention as the packages may block the workcell operation and may imply package loss or damage.

Results presented in Table 2 demonstrate that group T achieved significantly lower missed pick rates compared to group C. The statistical significance of this improvement is confirmed by the non-overlapping 95% confidence intervals between the two groups. The absolute missed pick rate reduction of 0.43 percentage points represents a substantial 19.25% relative improvement compared to the control group (C). The statistically significant improvement achieved implies significant benefits if deployed in real pro-

duction environments, where the scale of operations are even bigger than the considered testing environment. Additional failure modes exist in real production systems beyond missed picks. As shown in Table 3, we observed that there is no statistical significance between C and T1 in infeasible grasps rate, while T1 shows slightly higher multi-pick rates. These additional metrics indicate that the pick optimization does not improve the grasp feasibility, which is expected given the Pick Success Prediction (PSP) model was trained on predicting missed picks outcomes. On the other hand, we hypothesize that with improvements in the grasp qualities, the robot becomes more capable of picking up items, which can lead to higher chance of multi-pick.

Criteria for Model Selection: To deploy the approach on a large-scale robotic system, both software integration complexity and model capabilities need to be considered. There are straightforward software integration processes for the selected models, while maintaining sufficient model complexity for expressing picking success probabilities. As Table 1 shows, the gradient boosting model achieves superior RMSE performance across all 3 pick parameter dimensions compared to the 2-layer MLP, demonstrating higher accuracy in predicting optimal pick parameters. These findings led to the selection of the gradient boosting model for the evaluation of the proposed treatment against the control group. Its superior performance can be attributed to its ability to effectively handle categorical features. Its robust gradient boosting algorithm can capture complex relationships between input features and target pick parameters. The gradient boosting model’s performance establishes a strong baseline against which future, more sophisticated models -potentially employing deep neural network architectures - can be compared.

Directions Forward: While the current implementation has shown promising results, there are several opportunities for further enhancement. One limitation is the absence of rich visual information in the feature vectors. To address this, future work can explore machine learning architectures that efficiently process visual inputs, particularly through the use of image embeddings. This can enable the system to capture and utilize more comprehensive information about the manipulation environment. Furthermore, as the method is actively generating effective picks on the robot fleet, it provides valuable data on the physical world interactions. This data will be instrumental in training and evaluating more sophisticated pick success prediction models, which can subsequently further improve the system’s performance resulting in a lifelong improvement cycle.

Bibliography

- [1] Amazon Robotics: Robin (2022). URL <https://www.amazon.science/latest-news/robin-deals-with-a-world-where-things-are-changing-all-around-it>
- [2] Bohg, J., Morales, A., Asfour, T., Kragic, D.: Data-driven grasp synthesis—a survey. *IEEE TRO* **30**(2), 289–309 (2014)
- [3] Cao, H., Fang, H.S., Liu, W., Lu, C.: SuctionNet-1Billion: A large-scale benchmark for suction grasping. *IEEE RAL* **6**(4), 8718–8725 (2021)
- [4] Correll, N., Bekris, K.E., Berenson, D., Brock, O., Causo, A., Hauser, K., Okada, K., Rodriguez, A., Romano, J.M., Wurman, P.R.: Analysis and observations from the first amazon picking challenge. *IEEE TASE* **15**, 172–188 (2018)
- [5] Fang, H.S., Wang, C., Fang, H., Gou, M., Liu, J., Yan, H., Liu, W., Xie, Y., Lu, C.: AnyGrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE TRO* **39**(5), 3929–3945 (2023)
- [6] Fang, H.S., Wang, C., Gou, M., Lu, C.: GraspNet-1Billion: A large-scale benchmark for general object grasping. In: *CVPR*, pp. 11,441–11,450 (2020)
- [7] Lenz, I., Lee, H., Saxena, A.: Deep learning for detecting robotic grasps. In: *Robotics: Science and Systems (RSS)*, pp. 1–8. Berlin, Germany (2013)
- [8] Li, S., Keipour, A., Jamieson, K., Hudson, N., Swan, C., Bekris, K.: Demonstrating large-scale package manipulation via learned metrics of pick success. In: *Robotics: Science and Systems (RSS)* (2023)
- [9] Li, S., Keipour, A., Jamieson, K., Hudson, N., Zhao, S., Swan, C., Bekris, K.: Pick planning strategies for large-scale package manipulation. In: *Learning Meets Model-based Methods for Manipulation and Grasping Workshop, 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–4. Detroit, MI, USA (2023)
- [10] Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: *ICLR* (2023)
- [11] Liu, H., Deng, Y., Guo, D., Fang, B., Sun, F., et al.: An interactive perception method for warehouse automation in smart cities. *IEEE TII* **17**(2), 830–838 (2021)
- [12] Mahler, J., Matl, M., Liu, X., Li, A., Gealy, D., Goldberg, K.: Dex-Net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning. In: *ICRA*, pp. 5620–5627. Brisbane, Australia (2018)
- [13] Miller, A., Allen, P.K.: Graspit! a versatile simulator for robotic grasping. *IEEE RAM* **11**(4), 110–122 (2004)
- [14] Morales, A., Chinellato, E., Fagg, A., del Pobil, A.: Experimental prediction of the performance of grasp tasks from visual features. In: *IROS*, pp. 3423–3428 (2003)
- [15] Morrison, D., Corke, P., Leitner, J.: Learning robust, real-time, reactive robotic grasping. *IJRR* **39**(2-3), 183–201 (2020)
- [16] Wang, C., Vanbaar, J., Mitash, C., Li, S., Randle, D., Wang, W., Sontakke, S., Bekris, K., Katyal, K.: Demonstrating multi-suction item picking at scale via multi-modal learning pick success. In: *Robotics: Science & Systems (RSS)* (2025)
- [17] Yuan, W., Murali, A., Mousavian, A., Fox, D.: M2T2: Multi-task masked transformer for object-centric pick and place. In: *CoRL*, pp. 1–12. Atlanta, GA (2023)