Vulcan Pick: A Robotic System for Picking Targeted Objects from Fabric Pods

Kiru Park, Johannes Kulick, Alexander Melkozerov, Roc Arandes Vilagrasa, Teguh Santoso Lembono, Vanessa Neubauer, Artem Minichev, Kade Turner, Oana Agrigoroaiei, Pascal Klink, Jonathan Lee, Kshitij Dwivedi, Mazin Eltayeb, Ingmar Posner, Aaron Parness, and Can Erdogan

Amazon Robotics

***Abstract*—This paper presents an integrated robotic system designed for autonomous picking of targeted objects from cluttered and deformable shelves—a critical task in Amazon warehouse operations for processing customer orders. The system addresses common challenges in robotic picking including diverse object handling, densely packed storage, and dynamic inventories. However, shelf-picking introduces additional complexities, particularly the risk of inadvertently pulling adjacent items out, requiring advances in 3D scene understanding and adaptive motion control with continuous visual feedback. We introduce an end-to-end solution that combines proven classical methods with state-of-the-art approaches in computer vision, motion planning, and customized hardware. Our system has been operating in a live warehouse environment for over six months, processing more than 12,000 customer orders. This paper outlines the approach taken, presents key performance metrics, and discusses failure cases encountered. In highlighting the insights gained during this long-term deployment, and in particular the challenges in developing scalable robotic applications for warehouse automation, our aim is to communicate the current state of the art and propose future directions of development for robotic picking solutions.**

***Index Terms*—robotics, logistics, manipulation, picking, few-shot learning, object tracking, 3D scene understanding, robot vision, 3D perception, motion planning, grasping**

I. Introduction

In warehouse fulfillment operations, when a customer places an order, the corresponding item(s) needs to be picked from storage and delivered to packaging stations based on the quantity and type of purchased items. We focus on the picking process within modern goods-to-person warehouses where storage shelves (pods) are transported by autonomous mobile robots to dedicated picking stations. In these facilities, inventory is packed densely and heterogeneously. Once a pod arrives at the pick station, a person visually identifies the customer's product in a specific bin, extracts it, and places it into a container, which is sent to a packaging station once full.

This manual process requires precision, dexterity, and speed to ensure customer orders are fulfilled with the correct items on time and without defects. In this paper, we present an automation solution that handles up to 80% of customer requests while identifying robot-ineligible cases that are preemptively routed to the manual process. The system is designed to be deployed and operational 24/7 at warehouse scale.

K. Dwivedi was with Amazon Robotics when this work was performed.
I. Posner is with both Amazon Robotics and University of Oxford.

Fig. 1. Example of a fabric pod (left) and a number of its individual bins (right) containing a range of items. Elastic bands and bin lips at the bottom of the bins prevent items from falling out when the pod is transported to a picking station by a mobile robot. The metallic frame holds the fabric and obstructs items behind. Individual items are often obstructed or stacked. The images of the two bins (bottom-right) are captured while the band separation system, a subsystem of our overall autonomous pick solution, is holding the elastic bands during the extraction process.

Fabric pods are constructed from a heavy canvas material, partitioned into multiple bins mounted on 1-m-wide and 2.5-m-tall frames, as shown in Fig. 1. Each bin contains different products with a typical bin containing between two and eight items, some of which may be identical. Since storage space is a critical commodity for warehouses, the products are typically stored densely, in our case, requiring elastic retaining bands to prevent items from falling out during transport by the mobile robots.

Robotic manipulation in warehouses with everyday products requires dealing with a large variation in shape, weight, and texture of items, as well as considerations for deformability and fragility. The inventory in a single building spans

millions of unique products that are continuously updated to match customer demand; maintaining a priori information (e.g., visual models) becomes impractical. Instead, our system employs few-shot recognition using product images captured across warehouses, which is combined with metadata such as dimensions and weights, to identify targeted items without a barcode scan.

While items in well-organized bins—those with contents visible and unobstructed from the front—are easier to manipulate, the frequent picking and placement operations make it difficult to maintain such organization. Consequently, target items are often hidden behind or below clutter, requiring more complex motion strategies to search for the target items. We address this challenge through our *robot-eligibility* concept, which identifies such difficult cases in advance and routes them to the manual process.

The fabric pod design introduces additional unique challenges. The system must overcome obstacles such as elastic bands and *bin lips*—the bottom parts of the fabric bin frames. Unlike typical bin-picking tasks lifting items vertically, where gravity helps prevent items from falling, our horizontal extraction trajectory negates this benefit. In fact, gravity becomes problematic, creating significant momentum when items lose container support, often leading to drops. Finally, once an item is picked, placing it into a container safely and moving containers in and out of a workcell requires additional planning and perception capabilities.

This paper introduces an automated system, Vulcan Pick, that tackles these challenges using robot manipulators and ancillary mechatronic systems in concert with computer vision models and robot control algorithms to complete the target picking process. During a six-month deployment, the end-to-end integrated system handled more than 12,000 customer orders in an active warehouse, demonstrating over 90% success rate. In addition to introducing the overall system, this paper makes the following key contributions:

- Presentation of a fully featured perception and control system for extracting target objects from bins comprising: (1) a segmentation model that predicts both segmentation masks and spatial status of individual items, (2) efficient algorithms utilizing two 3D representations—signed distance function (SDF) and mesh—with segmentation labels for robust motion planning, and (3) continuous visual feedback enabling online motion adaptation and failure recovery.
- Development of an eligibility check framework that helps identify items eligible for robotic picking using existing imaging infrastructure, incorporating domain adaptation techniques and a depth prediction model that can see through semi-transparent elastic bands to bridge the gap between different imaging conditions.
- Introduction of specialized hardware components to utilize robotic manipulators for the picking and placement task including: (1) a movable conveyor system that receives items in front of target bins while preventing them from falling to the ground, and (2) three robotic end of arm tools (EoATs) designed to manipulate elastic bands, extract items, and transport items from the conveyor.
- Comprehensive evaluation of our system using data from a six-month field deployment in an active warehouse including detailed system performance metrics and analysis of common failure modes.
- Lessons learned from earlier explorations that motivated our proposed design choices, as well as next steps to advance our approaches through learned policies and a new hardware modality for improved capabilities.

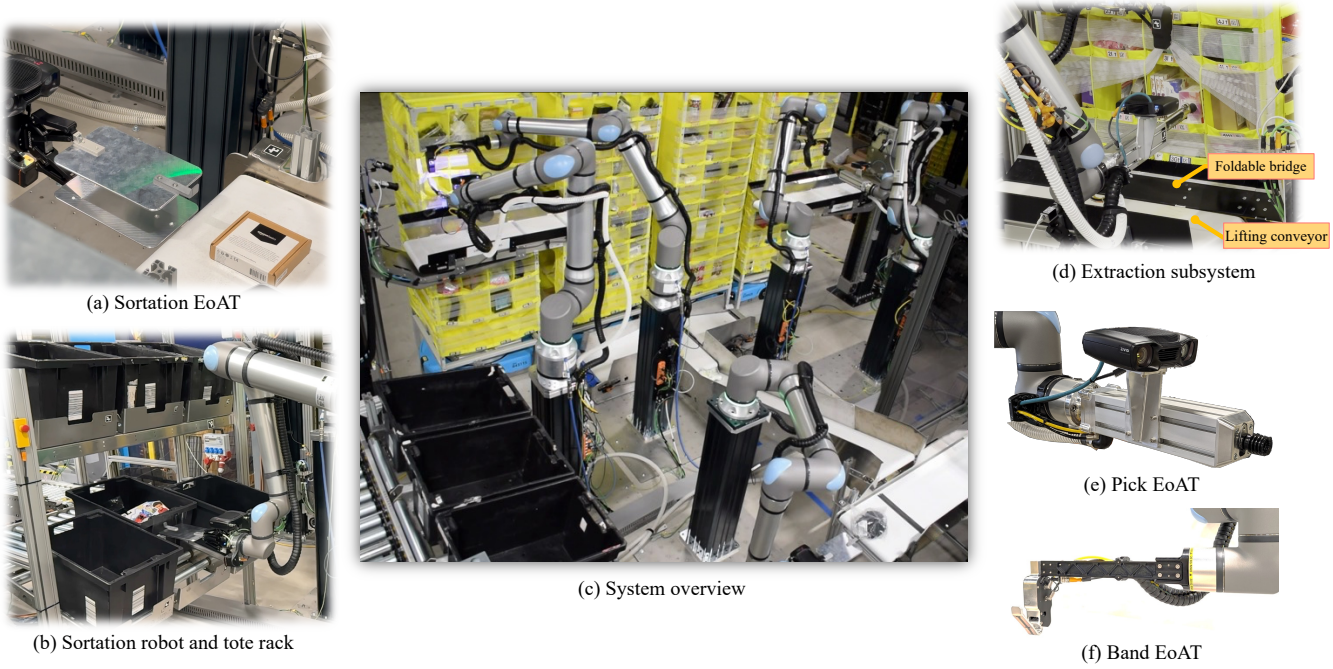
The remainder of this paper is organized as follows: Sec. II reviews related work in robot manipulation systems in logistics settings and their real world applications. Sec. III presents the problem statement and provides and overview of our proposed solution architecture. Sec. IV details our cell design and hardware components. Individual subsystems are described in Sections V, VI, VII, and VIII. We report deployment results and performance analysis in Sec. IX, share key lessons learned in Sec. X and suggest next steps in Sec. XI. We conclude in Sec. XII.

II. Related work

Several research initiatives have explored autonomous warehouse picking systems, with the Amazon Robotics Challenge (ARC) serving as a driving catalyst for advancement in robotic manipulation [1]–[3]. In the pick task during the last competition in 2017, 32 items were placed in a container, and the task was to pick and place 10 of the items into 3 cardboard boxes following a specific order in 15 minutes. 16 of the 32 items are sampled from a known set of 40 items that is provided to the participants before the competition, allowing them to train dedicated models to recognize those objects. The other 16 items are given to the participants only 45 minutes before each competition run to assess the generalization capabilities of the developed perception models. Under these conditions, the winning method achieved a 72% grasp success rate and spent 30 seconds on each pick attempt [3]. Apart from the limited knowledge on individual items, a subsequent challenge in the ARC picking task was to reliably grasp and transport target items. Multiple teams designed hybrid-type tools combining suction and fingered grippers to improve the success rates on certain item sets by switching modalities [2], [3], as they found reliable picking and transporting of diverse items with a single modality to be challenging.

Apart from the ARC, which is closest to the problem tackled in this paper, the RoboCup [4] and DARPA robotics challenges [5]–[8] have attracted researchers and engineers to build integrated systems for solving complex manipulation tasks in home and disaster environments. Although those robotics challenges were critical in catalysing problem understanding and to focus research effort, the systems proposed at the time did not consider scalability and long-term operation with minimal operator intervention. Moreover, their artificial environments did not fully represent real-world deployment challenges, particularly in terms of object diversity and system reliability requirements for longer-term daily operations.

Outside of academic research, there are several robotic applications deployed in logistics where packages or items are picked from conveyors, containers, or pallets [9]–[14].



(a) Sortation EoAT

(b) Sortation robot and tote rack

(c) System overview

(d) Extraction subsystem

(e) Pick EoAT

(f) Band EoAT

Fig. 2. Overview of hardware components and the cell design. (a) The sortation End of Arm Tool (EoAT) transporting extracted items from the conveyor to totes. (b) Sortation robot and tote racks receiving empty totes and sending full totes to the next processes. (c) Overview of the system having two extraction systems paired with a sortation system. (d) Extraction system consists of two robots and a movable conveyor, lifting conveyor, (e) *Pick EoAT* used to extract items from bins, (f) Band EoAT used to hook bands and hold bands while extracting items.

These tasks are typical in warehouses and are widely solved using suction-based tools. In this scenario, objects are lifted vertically off the ground, requiring the generated suction pressure to overcome the force of gravity. Due to gravity, it is unlikely to extract other items in a container as long as the suction tool is engaged on the correct item. For shelf-style storage systems where items are extracted horizontally, this beneficial effect of gravity does not apply.

Outside of warehouses, there exist efforts to pick orders from supermarket shelves [15] and datasets [16] to initiate training of perception models for the diverse object sets presented in supermarkets. Since supermarkets organize and group similar items together, target items are rarely occluded by clutter. Therefore, robots are mostly able to pick items at the front-most locations, easing the manipulation task. The situation is different in our scenario, where different items can be randomly placed inside bins and other items or obstacles presented in bins can often occlude the target item.

This paper introduces an integrated system for picking items from cluttered bins in large-scale everyday operations. While the ARC highlighted key manipulation challenges, our system extends beyond these to address real warehouse complexities: handling millions of different items in cluttered bins and integrating seamlessly with existing infrastructure. Through a six-month deployment, the system achieved more than 90% success rate, demonstrating its effectiveness as an end-to-end automation solution. The next section details the specific challenges and tasks our system addresses, along with its high-level workflow.

III. Problem Statement and Solution Architecture

The goal of the pick process is to pick a target item from a target bin and deliver it to a specified box-like container, a tote (the black boxes in Fig. 2-b&c), which accumulates multiple items to be shipped to downstream processes such as packaging. As shown in Fig. 1, the items are stored in the shelf-style container, fabric pods. the fabric pods are made of a fabric material that is partitioned into multiple bins. Each bin has a unique ID and the warehouse system tracks a list of items in the bins so that requests for picking specific target items from them can be generated.

Due to varying item sizes and weights as well as the endless combination possibilities of individual items in a bin, certain pick requests are highly challenging. Reasons can range from different items of almost identical appearance within a bin to item configurations that are beyond the capabilities of our hardware embodiment, e.g. by providing no free surface for engagement with the suction gripper due to occlusion. Thus, we introduce the *robot-eligibility* concept, which distinguishes items that are both identifiable and pickable from those that are not. Pick requests for non-eligible items are assigned to manual stations whereas eligible items are processed by our solution.

The overview of the pick workflow is visualized in Fig. 3. The system can be divided into three subsystems: Eligibility check system, Extraction system, and Sortation system. Fig. 3 highlights key steps in the three subsystems. The system is integrated into an active warehouse to accept pick requests of eligible items from real customer orders through communica-

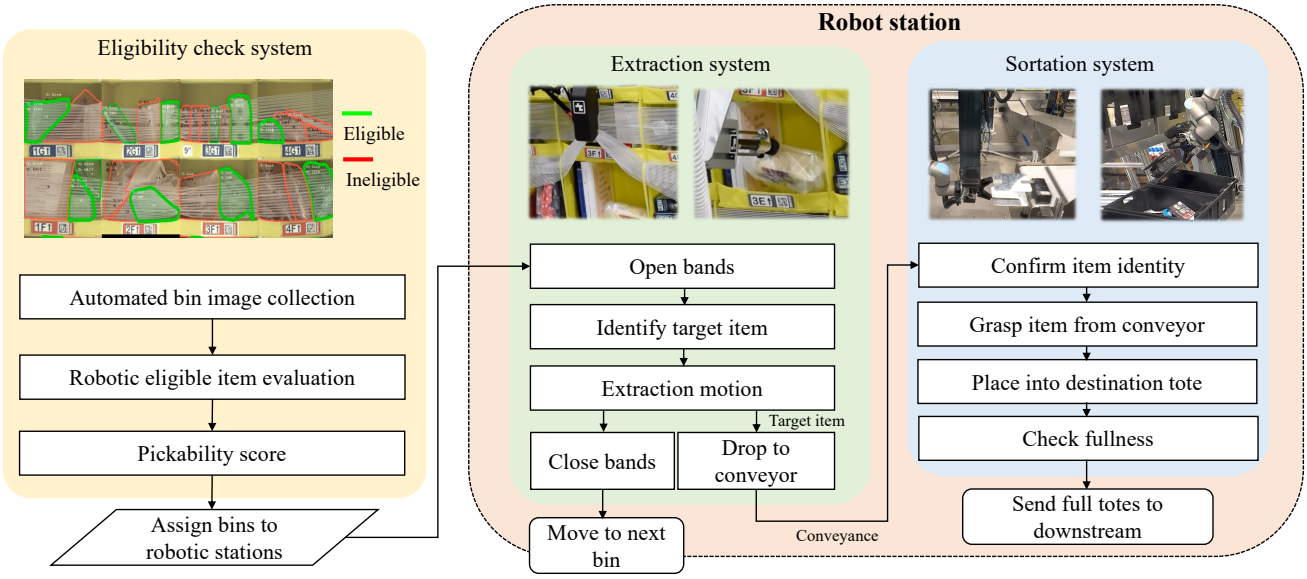


Fig. 3. Overview of the end to end process running in the cloud infrastructure (the eligibility check system), and within the robot station (the extraction and the sortation systems). The eligibility check system is running in a cloud environment to track status of bins using images captured from data collection tower existing in fulfillment centers. The system informs which items can be assigned to the automated picking station. When a pod arrives at the station, the extraction system opens bands, and extract target items from target bins. As soon as a target item is dropped to the conveyor, the sortation system transports items to a destination tote through conveyance and a sortation robot. When a tote is full with items, sortation robot pushes the tote to send it to downstream paths.

tion with the software infrastructure of the warehouse system. The sortation system is connected to the existing outbound conveyor system to deliver items to the downstream processes.

A. Eligibility check

The eligibility check system is running in the cloud to process image data collected from existing data collection infrastructure, so-called visual bin inspection (VBI) towers. The tower captures images of bins to track their current content. A bin image comes with metadata about items in the bin such as quantity, item identifier, title, weight, and dimensions of each item. The eligibility check system processes both images and metadata to decide if each item is eligible to be picked at the robotic station or not. The left block in Fig. 3 visualizes example results where items outlined in green indicate robot-eligible items and red indicate robot-ineligible items. We describe the algorithms and criteria of the eligibility check in Sec. VIII. The eligibility results are shared with the warehouse software system so that pick requests for robot-eligible items are assigned to Vulcan Pick stations.

B. Extraction procedure

When a pick request is assigned to the station, the pod containing the target bin is delivered to the station by a mobile robot. The extraction system detects the target item within the target bin, picks the item from it, and drops it onto a conveyor (Figure 2-d). As elastic bands cover the bin to prevent items from falling out, the system initially opens the bands to obtain a clear sight of bin content and extract items without blockage. When the item is successfully extracted, the system releases the bands, moves both robots to states that allow the pod to leave the station safely, and continues to the next bin.

C. Sortation procedure

The first conveyor making contact with the item (Figure 2-d) can be lifted up and down in order to safely catch and transport the item from various heights to a chain of conveyors. The conveyor chain transports the item to the so-called sortation robot (Figure 2-a+b), which places the item into the destination tote while tracking the fullness of individual totes to replace them with empty ones if required.

IV. Cell Design and Components

Figure 2-(c) shows an overview of the station and hardware components. The system comprises five robotic manipulators and several mechatronic systems. We used robotic arms mounted to fixed locations for tasks requiring flexible motions to handle items and containers. The mechatronic systems are used to bridge between robots to transport items or move target containers efficiently to complement the limited workspace of the fixed robot arms. One workcell includes two extraction systems paired with one sortation system. Fitting two extraction systems into one workcell reduces speed requirements of the individual extraction system by a half. This allows the extraction system to interact with items longer in challenging cases and reduces defects.

A. Extraction system

The extraction system (Fig. 2-d) performs two physical actions: band separation and item extraction. We employ two robotic arms, the *band robot* and the *pick robot*, that are equipped with dedicated EoATs, the band EoAT and the pick EoAT, specialized to tackle these tasks. We use robot arms for both band separation and item extraction tasks. The robot arms

can handle up to 20 kg payload and come with a 1,750 mm reachability, which is sufficient to reach pods having heights of up to 2,574 mm. Their payload limit is also sufficient to handle target items of up to 2.27 kg and holding elastic bands in addition to carrying the weights of the EoATs with a margin. The robot arms are equipped with a force/torque sensor (F/T sensor) at the tool interface measuring forces and torques applied to the EoATs. Both robots utilize the F/T sensors to perform force-based control and detect abnormally high forces applied to the EoATs to prevent potential damages and failures. We present the extraction system in sections V and VI.

B. Eye-in-Hand Camera

Visual information is critical for performing band separation, motion planning, item identification, and extraction behaviors discussed in sections V and VI. In the early stage of system development, we explored both externally mounted, static cameras and a camera mounted on the wrist of the Pick robot. We identified that the camera mounted on the arm provided more flexibility in terms of camera poses when capturing images of the target bin. We control the robot to capture target bin images from a constant relative pose with respect to bin centers. The arm-mounted cameras also provide continuous obstruction-free image observations while the robot is interacting with items in bins. Therefore, we mounted a camera, the *pick camera*, on the pick EoAT to acquire images of individual bins and monitor robot-item interaction. We use an industrial RGB-D camera that produces RGB and pointcloud data using structured light projection. We use two capture settings: one that acquires high quality images for core perception tasks producing 1224 x 1024 px resolution at 500ms acquisition time, and another to acquire continuous visual feedback during item interaction at a 10Hz acquisition frequency with a lower resolution of 612 x 512 px.

C. Lifting Conveyor

Gravity introduces challenging dynamics when picking objects from bins as soon as the objects are no longer supported by the bottom surface of the bins. Furthermore, transporting objects from source bins to destination totes using a single robot is inefficient as the robot has to move back and forth between the bin and the drop-off location. To tackle these challenges, we developed a so-called *lifting conveyor* that receives objects right below the target bins and transports them to the next conveyance (See Fig. 2-c&d, the conveyors in front of pods). We found this system to improve the extraction cycle time while preventing unnecessary defects such as dropping extracted objects to the ground floor, which often causes severe damages to items. The lifting conveyor consists of a conveyor that is moved vertically by a linear actuator and an active bridge that is actuated by pneumatics (the black plate in front of the pod in Fig. 2-d). The active bridge fills the gap between a pod and the conveyor by folding and unfolding a metallic plate while extracting an item from a bin. The bridge is folded before and after an extraction to secure items on the conveyor while moving the conveyor down and provide sufficient margin when a pod enters and exits the station.

D. Conveyance and Sortation robot

Objects from the lifting conveyor are transported via multiple conveyor segments to the sortation robot (Figure 2-a). This robot is equipped with a hand-mounted camera that serves dual purposes: estimating the dimensions of incoming items on the conveyor to optimize grasp planning and placement strategies, and monitoring empty space in the destination totes by capturing images to determine where to place items. The Sortation robot grasps incoming objects and deposits them into one of six destination totes. The details of the sortation system are presented in Section VII.

E. Tote racks and tote handling mechanism

Totes serve as the final destinations for items retrieved by our system. These totes must be transported to subsequent process paths when full, while empty totes need to be brought in to continue picking operations. To automate this process, we utilize an existing mechatronic system, the Amazon Robotics Semi-Automated Workstation (ARSAW) [17], which extracts totes from a stack and supplies them to a rollable conveyor rail. When an empty tote arrives, the Sortation robot grasps it and places it in an empty slot in the rack. When a tote is full, the Sortation robot pushes it onto the rollable conveyor rail, from where the ARSAW transports it to the outbound conveyor system.

F. Visual Bin Inspection Tower

In Amazon robotic warehouses, VBI towers [18] (automated imaging stations) are installed throughout the warehouse to capture bin images, enabling continuous tracking of bin status and validation of inventory accuracy. The pods are placed in front of the VBI towers after they visit stations close to a tower. While they are not a part of our cell design, images from VBI towers are used to estimate item eligibility for our robotic picking station. We discuss the eligibility check process using monocular images from VBI towers in Sec. VIII.

The following sections describe details of EoATs and algorithms used in the extraction system. The extraction process consists of two main components: band separation and item extraction. Fig. 4 illustrates the perception and control workflows for both components.

V. Band Separation

The band separation process aims to clear obstructing bands from the target bin, enabling *the pick robot* to access and extract the target item. To accomplish this, we use a hook-shaped EoAT to lift the bands from bottom to top. This process requires significant force—up to $70\,\mathrm{N}$—applied through the hook. This motivates us to minimize interaction with items behind the bands to prevent potential damage on the items. To achieve precise band manipulation, the system first captures an image using *the pick camera* mounted on *the pick robot* to estimate both band locations and gaps between bands and items. This spatial information enables *the band robot* to safely position the hook behind bands and move the bands using force-guided control.

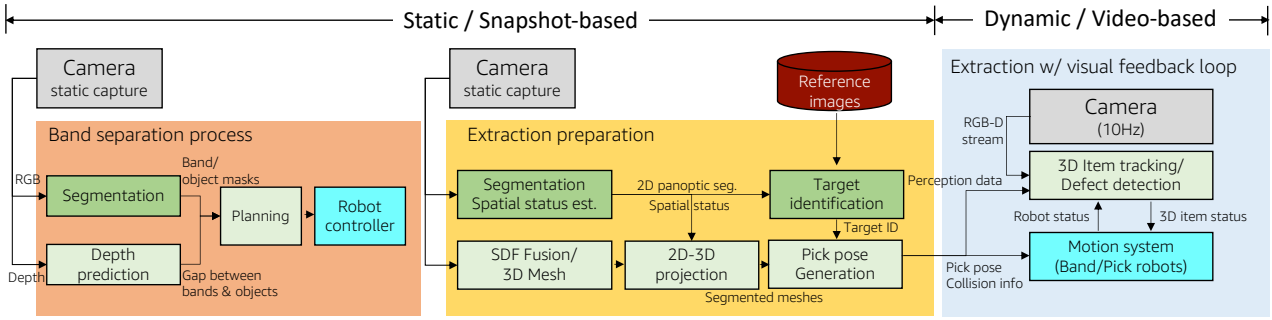


Fig. 4. Overview of the core perception pipelines during band separation process, extraction preparation, and extraction motions. The first two stages uses static images captured at the beginning of each process to generate initial trajectories. In contrast, the camera is continuously triggered during extraction motion at 10Hz to track items in 3D and adapt motion strategies and trajectories accordingly.

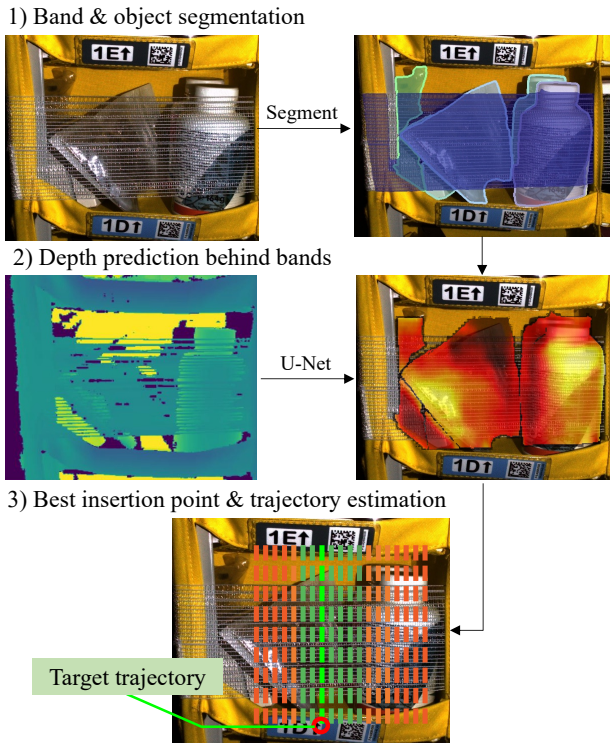


Fig. 5. Perception and motion planning process for band separation. Both band and instance masks of objects are predicted using a segmentation model [19]. Both RGB and depth images are fed to a U-Net network to predict gaps between bands and items behind bands (brightness indicates closeness to the bands). The planning algorithm evaluates straight upward trajectories in each vertical slot to find safest location and depth of the hook to minimize interaction with items behind bands to avoid damages. The system identifies multiple valid trajectories (shown in green lines) and evaluates risks of each trajectory to find the optimal path (highlighted in bright green) and the contact point to insert the hook (the red circle).

1) Band EoAT: The band separation subsystem features a custom-designed *band EoAT* with an active hook mounted on *the band robot* (see Fig. 2-f). The active controllable hook provides an additional rotational degree of freedom (DoF) and system compliance through a double-acting pneumatic cylinder. This cylinder actuates the hook via a lever mechanism, enabling a 30-degree range of motion. The lever mechanism's profile is specifically designed to ensure robust band grasping even when bands are partially overlapped with the hook during closure. The EoAT's L-shaped offset design optimizes *the band robot*'s reachability while minimizing the tool's volume to avoid limiting *the pick robot*'s workspace due to potential collisions.

2) Band Separation Planning & Motion: As previously explained, we leverage the *pick camera* attached to *pick EoAT* to support band separation planning and validation. The band separation workflow operates as follows: The pick robot positions itself at the target bin to capture an RGB-D image. Based on this perception data, the system computes an optimal hook insertion point and generates a corresponding motion profile, which the band robot then executes. Upon detecting any failures, the system initiates reattempt procedures until either achieving successful separation or reaching the maximum attempt threshold. When the system fails to open the bands, it releases the pick request and assigns it to the manual stations.

Fig. 5 illustrates the band separation process and intermediate results. Given both the segmentation masks and the depth data, we compute the optimal insertion point (shown as red circle) that minimizes the chance of interaction between the EoAT and the items inside the bin during band separation. The vertical insertion location is determined based on the height of bin lips whereas the horizontal location is chosen to be on the vertical slot with the lowest cost. A slot will have low cost if there is no item close to the bands within the slot.

VI. Item Extraction

The perception and control architecture of the item extraction step is visualized in the second and third step in Fig. 4. When the *band robot* finishes its motion, the *pick camera* captures an image to confirm if the bands are cleared from potential item extraction paths. Once successful band separation is confirmed, the extraction preparation workflow is triggered to build a plan to extract a target item from the bin. This section introduces detailed specifications of the EoAT used to pick items and the core algorithms that process the perception data and generate the robot motions.

A. Pick EoAT

The *pick EoAT* uses a suction cup to extract items (see Fig. 2-e). The *pick EoAT* features an active linear extension mechanism that provides an additional DoF with a 200 mm

travel range to reach deeper into the bin when extended. The EoAT employs a double-acting pneumatic cylinder that drives the suction cup extension and retraction. An electro-pneumatic pressure regulator enables remote control of the cylinder pressure, allowing dynamic adjustment of the extension DoF compliance for optimal interaction with target items. The EoAT integrates comprehensive sensing capabilities through a vacuum pressure sensor at the tool manifold and a flow meter mounted at the robot pedestal, providing real-time feedback on suction cup engagement quality.

B. Core perception for 3D scene understanding

To generate safe robot motions for successfully extracting items, it is essential to know the 3D locations of and spatial relationships between individual objects in a bin as well as the location of bin structures itself. Object locations and their boundaries are necessary to locate suction-based grasp poses that are close to object centers and have minimal interaction with neighboring objects. The bins consist of deformable structures (e.g., the side walls, ceiling, bin lips, and elastic bands) and rigid structures (e.g., the metal bar structure at the side most bins). Therefore, it is beneficial to recognize these structures separately to inform the motion planner about compliance around deformable areas and avoid collision with rigid structures.

The goals of the core perception pipeline are: (1) to provide information about which parts of the 3D space are occupied by objects, flexible bin structures, or rigid structures, (2) to detect planar areas safe for suction engagement, and (3) to understand spatial object-object and object–bin relationships to adapt our motion strategy. We use a segmentation model with a spatial relationship head to recognize objects and bin structures in the 2D image space and project the results onto 3D SDF grids and meshes to achieve the goals.

1) Segmentation with a spatial relationship head: We use a customized version of MaskDINO [19] with a Resnet-50 backbone [20] to produce instance-level segmentation masks on (1) object instances, (2) bands, (3) metallic bars, (4) fabric areas, and (5) name tags of each bin. We collected images with and without bands on the bins and manually annotated segmentation masks on more than 200K bin images to train this model. It is worth mentioning that our early prototype used MaskRCNN [21], and showed artifacts when multiple items were slanted together. We realized that this problem is caused by ROI alignment and resizing process that amplifies prediction errors. In contrast, modern segmentation architectures that use tokenized object representations such as Mask2Former [22], [23] and MaskDINO predict masks directly on the full image space, reducing the impact of noise on the final prediction. In addition to the standard segmentation head of the MaskDINO model, we added an additional classification head at the last layer to predict a 4-dimensional vector representing the spatial status of each instance. Each vector value corresponds to the probability of the segment being (1) not an item (e.g., bands or metal bars), (2) an item in good status (not obstructed), (3) an item below others, or (4) an item blocked by others. Fig. 6-a) visualizes the segmentation and status prediction of our model.

2) SDF and Mesh Representation: In addition to the semantic understanding obtained from RGB data, we process pointcloud data from the camera to represent 3D geometries of items and bins. We reconstruct a signed distance function (SDF) [24] as a representation of the 3D scene. The depth map from the camera updates the SDF grid along rays of each depth pixel. The grid represents occupied volumes behind observed surfaces as negative values and unoccupied volumes closer to the camera as positive values. The raw SDF values are processed by a 3D Gaussian filter to reduce noise from the raw pointcloud. We then build a mesh of the scene via the Marching Cube algorithm [25] (Fig. 6-c). The segmentation masks are projected to the scene mesh to split the mesh into smaller meshes representing individual items. In addition, the segment labels are projected to the SDF grid close to surfaces of the mesh in order to represent which part of space is occupied by which item in the uniform 3D grid. The labeled meshes and the labeled SDF grid provide comprehensive knowledge of item surfaces as well as occupancy of space within the bin to inform pick pose generation and motion planning.

C. Item identification

The item identification (ID) module identifies a target item within the current bin using the item's visual appearances. The system has access to reference images of all items in the current bin captured from different imaging mechanisms across warehouses. The ID module then has to identify target item segments from the current bin using the reference images. The described identification module has to address two challenges:

- Reference images not available during training: Given the scale of Amazon catalog (millions of items) and its continuously changing nature, we can only train models on a limited set and need to ensure that the models generalize to previously unseen reference images.
- Different lighting and resolution between inference and reference images (Figure 6-b): The camera types and capture settings between our robotic station and the imaging systems that produce reference images differ, resulting in a different appearance of in-bin and reference image.

To address these challenges, we train a model which maps RGB images into an embedding space such that images of same item from different image sources are closer in embedding space than images of other items. We take inspiration from MoCoV2 [26] and adapt it to our use case. In our setup, positive examples are images of the same item from different sources while negative examples are images of other items. In this way, the model learns to map the images of the item from different sources closer in embedding space while discriminating images of other items. We also considered DINO [27], which only requires positive examples, and has been shown to outperform MoCoV2 [26] on several benchmarks. However, in our case, we found that training with negative examples leads to a clearer decision boundary between correct and incorrect matches.

Given N segments containing items in a bin and reference images of M items known to be in the bin, our goal is to match each item with its best corresponding segment based on distances in the embedding space. We create a $N \times M$ distance matrix which contains distances between segments and reference images in the embedding space. Then, we apply the Hungarian matching algorithm [28] to solve the corresponding assignment problem.

D. Pick pose generation

The suction-based *pick EoAT* requires stable engagement of the EoAT's tip on object surfaces. An important factor for generating strong engagement is to ensure uniform contact between the opening area of the suction cup and the object surface. This uniform contact can be established by aligning the EoAT along the surface normal of a planar area on the object surface. Based on this idea, we exploit the mesh representation of objects to detect planar areas on their surfaces. We then generate pick samples on these planar areas and filter and rank them according to different criteria to identify promising pick poses. The individual steps of this pipeline are explained in the following sections.

1) Planar patch detection: We adapted classic region growing algorithms [29]–[31] with known connections of vertices to sample patches from uniformly sampled vertices, and expand the patches if their adjacent vertices have similar surface normals (Fig. 6-d). A patch stops expanding if it becomes bigger than a threshold, or there is no more vertex having similar surface normals. As a result, we can represent object surfaces with planar patches, and each patch has a representative normal at its centroid.

2) Sampling pick candidates with skewing: After patch detection, we create initial pick candidates that make contact at the centers of planar patches and are aligned with the patch normals. However, due to occlusions by the bin structure and objects within the bin, some pick poses are not accessible without collision. At the same time, the suction cup has a soft material that can deform its shape. This deformation enables good suction engagement even with slightly skewed approach angles. Thus, we generate additional pick candidates from the initial candidate by applying different rotation angles up to 15 degrees. As a result, we generate multiple pick poses with the same contact point but varying approach angles. The dark green lines in Fig. 6-d) represent multiple pick candidates created by this process.

3) Filtering and scoring: Depending on the sizes of object surfaces, the previously described process generates an average of 1,000 pick candidates per object. Our strategy in the pick generation process is to generate as many candidates as possible to maximize the likelihood of finding at least one good pick pose. However, this approach requires us to quickly filter out infeasible or low-quality pick poses to maintain high throughput. The SDF representation with object labels plays a significant role in this strategy:

- Fast collision checking: we use the SDF to perform fast collision checking and remove pick poses in collision. We represent *pick EoAT* as points using vertices of the 3D CAD model, and transform them to a candidate pick pose. We check if there is any point inside occupied volumes via the SDF values. The process requires reading values from a 3D array for each point instead of using a search algorithm to detect nearby objects, which is computationally more efficient.
- Detect suction contacts: the object labels within the SDF enable us to derive information around the suction cup including average distance to adjacent object or bin structure surfaces. We again represent the 3D suction cup volume as 100 points and check the SDF values and object labels of the grid cells occupied by those points. In this way, we can easily evaluate if the suction cup has uniform contact on the target without contacting on other items.

We filter out pick poses that are risky to execute due to collisions or a high probability of touching nearby objects. The remaining pick poses are evaluated based on the quality of uniform suction contact around the suction cup area and their surface normal alignments. Higher scores are given to pick poses that are close to the centroids of the objects as they prevent unnecessary rotation of objects which can lead to the detachment of suction during robot motions.

Overall, the hand-designed evaluation criteria prioritizes pick poses that are most likely to be successful. It is worth mentioning that the pick pose generation pipeline can be further improved given that the SDF grid can be directly used to compute the gradient of pick poses to improve uniform contacts, and adjust pick poses to avoid collisions. Furthermore, SDF grids with object labels can be used as an input to 3D neural networks such as 3D-CNNs to realize a learned pick scoring approach, as has been explored for similar suction-based EoATs [32].

E. Motion with proprioceptive feedback

The motion planner module accepts a set of feasible pick poses from the previous step and generates a feasible motion plan to pick the target item. The motion plan is classified into two parts, i.e, free space and task space motion. During free space motion, the robot moves to the target bin and approaches the target item to engage suction. The item is then extracted from the bin as part of the task space motion.

The free space motion is defined by a set of waypoints generated based on the pick pose. Given a pick pose, the planner generates approach waypoints at pre-defined offsets from the pick pose and returns the highest-scored pick pose whose waypoints are all feasible, i.e., can be reached without collisions or kinematic singularities.

The waypoints are then given to the robot controller which tracks the waypoints with linear interpolation in task space.

The task space motion consists of a set of movement primitives designed to extract the target item from the bin while overcoming two obstructions: the bin lip and the metal bar. After engaging suction on the target item, the robot moves the item to the center of the bin in order to avoid the metal bar and the bin lip corners. Then, it lifts the item above the bin lip and finally extracts the item out of the bin. Throughout the

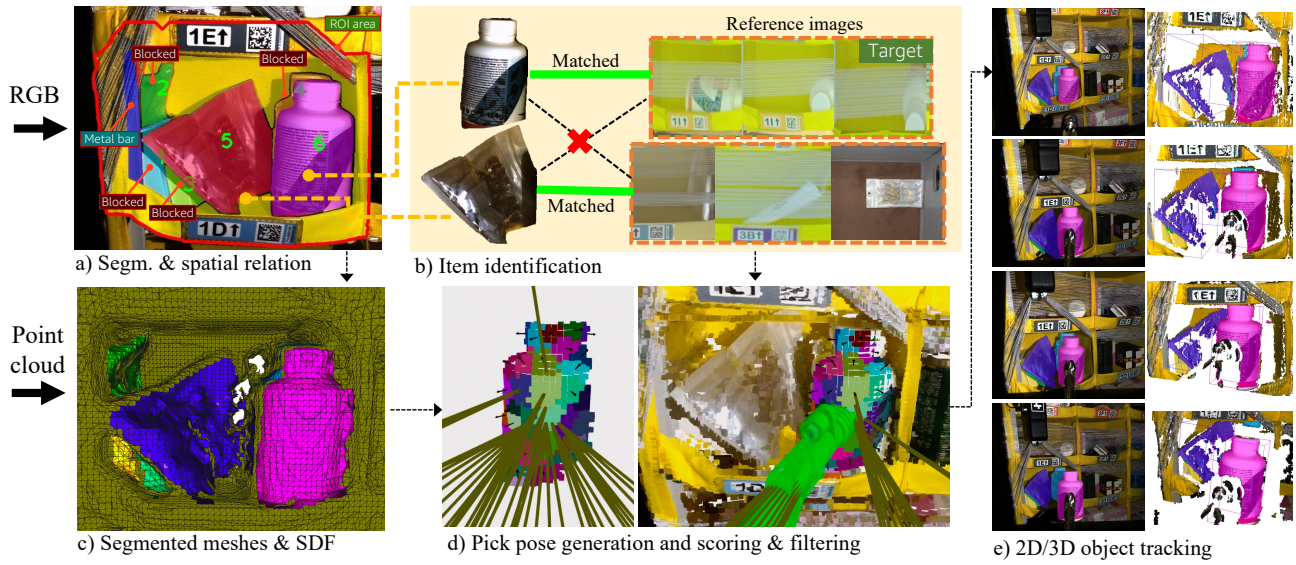


Fig. 6. The visualization of perception data during the extraction process. a) Results of segmentation model [19] with a spatial estimation head. The segmentation model predicts instance masks of individual objects as well as masks of metalic bar, bin label, the blue box with '1D', and yellow fabric area from a target bin. b) Item identification uses reference images collected from other processes in warehouses to detect a target item, c) Pointcloud data is processed to Signed Distance Function (SDF) and meshes. The segmentation results in 2D are projected to both SDF grids and segmented meshes. d) Pick pose generation process splits surfaces of a target item into small planar patches to create initial pick poses. Pick poses are augmented to allow variations of approach directions, and evaluated to rank them using SDF values around contacting area of the suction cup and collision. e) Visualization of 2D and 3D item tracking results of all objects in the bin while the target is extracted from the bin.

motion, we monitor the force/torque and pressure sensors to detect a potential loss of suction and avoid applying excessive forces on the bin surroundings. Additionally, we monitor the RGB-D image from the pick camera during extraction, which will be elaborated further in the next section.

F. Continuous visual feedback and 3D object tracking

In addition to recovering from suction losses during the task space motion, our system also needs to cope with the risk of co-extracting non-target items. This risk is higher compared to vertical picking since gravity acts perpendicular to the extraction direction of items, thereby not pulling items back into the container.

To correct these potential failures during the task space motion, the camera continuously captures images to track 3D poses of all objects in the bin. The camera on the *pick EOAT* acquires RGB and pointcloud data at 10Hz. The initial segmentation mask from the core perception pipeline is used to initialize Cutie [33], a state-of-the-art video segmentation tracking approach. We then project the tracked segmentation masks onto the pointcloud to locate each object in 3D as shown in Fig. 6-e. We align each point-cloud in the robot's coordinate frame using the respective robot pose at capture time.

In contrast to open-loop robot motions, visual tracking of 3D object locations unlocks adaptive robot behaviors, thereby improving the success rate and the quality of robot-object interactions. We implemented three behaviors which leverage the continuous visual feedback instead of going through the core perception pipeline from scratch:

- Re-attempt for stronger suction engagement: We measure if an initial contact on a target surface is successful via the vacuum pressure. If the pressure is low due to an improper suction engagement, we re-generate a new pick pose on the object surface using the tracked item locations. Even though the initial contact often pushes items away or to a side, visual tracking enables the generation of a pick pose on the target item at its new location. We limited the number of reattempts to three to avoid spending too much time on failed cases.
- Re-attempt to recover from failures: Even after proper suction engagement, obstacles along the extraction trajectories such as the bin lip, the metal bar, and non-target items can result in a loss of suction during motion. Thus, when the suction pressure indicates that the target item is no longer attached to the *pick EoAT*, the robot uses the updated item pose to generate a new pick pose and attempt picking the item again.
- Failure monitoring and prevention: In contrast to the first two use cases where suction pressure primarily triggers actions, this use case is triggered by the visual tracking pipeline. The tracking pipeline monitors the 3D locations of items with respect to the front face of the bin. When non-targeted items are about to move outside the bin, the pipeline triggers a failure prevention behavior that puts the items back into the bin and avoids picking wrong or multiple items.

It is worth mentioning that our initial baseline used optical flow to track point-level correspondences using RAFT [34]. However, this baseline showed limitations when new object surfaces are exposed during robot-item interaction as it is not designed to associate novel pixels that are not observed from previous frames. In contrast, Cutie [33] is capable of reliably propagating segmentation masks to unseen surfaces of items

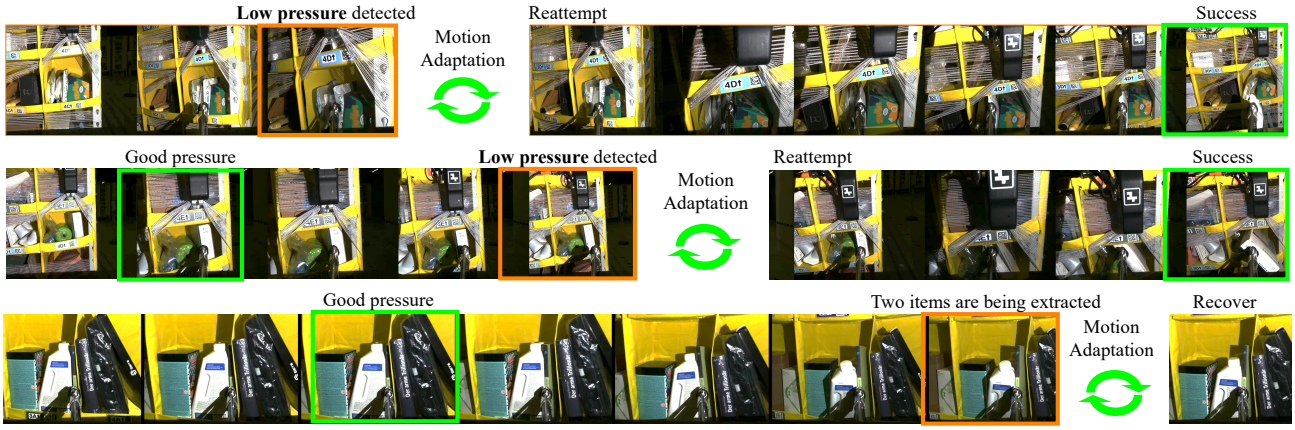


Fig. 7. Three uses cases of the visual feedback with object tracking. First row: The robot recomputes the target pick pose if the initial pick pose failed to engage suction on the target item using latest location of the target item. Second row: the initial suction was engaged but blocked by the bin lip (the bumper at the bottom). The robot recomputes a new pick pose to pick the target again using the latest location of the target item. Third row: The robot was extracting the target (the book) but pulling the neighboring item. The robot pauses the motion and put items back to avoid picking the unintended item.

while being robust against occlusion. We therefore used this method to track item segmentations in 2D image space.

VII. Sortation

The sortation system receives items through conveyors connecting the extraction system to the sortation robot (see. Fig.2-c). The target item extracted from the target bin is dropped to the lifting conveyor, and the lifting conveyor passes the item to consecutive conveyors, and the sortation robot grasps the item from the conveyor to drop it to the destination tote. The sortation robot checks if the tote is full or not and pushes the full tote to the rollable rail to send it to the downstream process. The sortation robot pulls a new empty tote to the rack to continue.

A. Sortation EoAT

There are two primary tasks performed by the sortation robot: (1) grasp an item and place the item into the tote, and (2) push and pull totes. To achieve these tasks reliably, we use a commercial parallel jaw gripper with custom fingers that support items from top and bottom as shown in Fig. 2-a. The finger plate has a hook that is used to pull empty totes into the inclined rack using the robot. Similar to *pick EoAT*, a RGB-D camera is mounted on the *Sortation EoAT* to estimate sizes and locations of items on the conveyor before grasping, estimate fullness of the tote, and optimize dropping locations to place items safely.

B. Grasping item from conveyor

To grasp an item from the conveyor, *Sortation EoAT* is located at the edge of the conveyor while the conveyor pushes the item to the EoAT. We capture an RGB-D image of the item on the conveyor to segment and estimate sizes and orientations of the item. We trained a segmentation model [19] to segment items from conveyors and fit a 3D bounding box to estimate the size and orientation of the item. The height of the item informs the height of the top finger plate while the bottom plate is aligned slightly lower than the conveyor. The location of the bounding box informs the lateral location of the item so that the item can be placed at the center of the finger plates.

C. Fullness estimation and placement

After dropping an item in a destination tote, the camera captures an RGB-D image of the tote to estimate empty spaces available within the tote. The space information is used to decide where to place the next item. The pointcloud is processed as a 3D occupancy grid in 1 cm resolution. It assumes spaces behind observed surfaces are occupied by items through the bottom of the tote. The pointcloud of a new item collected during the grasping process is also converted to voxels in the same resolution. By using a convolution between 3D voxels representing the environment and the item, we compute where the item can fit in the container. A target position is then chosen by maximizing the distance to the closest occupied voxel. This also ensures an approximately uniform filling of the tote. When the item size is bigger than the free space, the robot replaces the tote with a new empty tote.

VIII. Eligibility Check

VBI towers discussed in Sec. IV capture images of bins frequently. These bin images allow us to inspect and verify multiple conditions: whether the item identification model can detect target items from the bins, whether items are obstructed by other items or obstacles, and whether items can be reliably picked with the pick EoAT. However, VBI tower is equipped with three monocular RGB cameras without a depth camera and bins are covered by bands (see Fig. 8 for an example). Since the core perception pipeline uses RGB-D data captured after opening bands, we cannot directly apply our core perception pipeline due to domain gaps between the two different image acquisition settings and modalities. This section discusses our approaches to adapt domains by training dedicated models on the new domain and using domain adaptation approaches to transfer knowledge from a

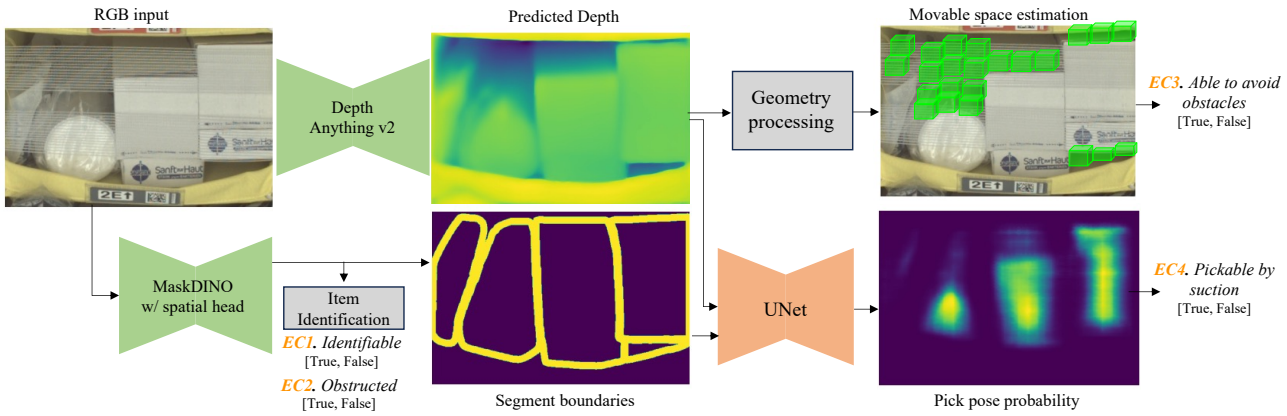


Fig. 8. Overview of the eligibility check pipeline filtering target items ineligible for the system. We use existing data collection towers to decide eligibility. The same segmentation model [19] with a spatial status estimation head is fine-tuned on this domain, and the same item identification model is used to identify each item in the bin. This process determines if each item is identifiable and whether it is obstructed. We fine-tuned Depth Anything model [35] to estimate geometries of objects and free space within the bins. This determines whether each item has enough free space to be lifted over bin lips and moved to the side to avoid metal bars. We transfer generated pick pose information to this domain to train a U-Net estimating pick pose probability, ensuring that pick poses can be generated on the item surfaces. When an item satisfies all eligibility criteria, the item is marked as eligible, and can be assigned to the robotic stations.

source domain with ground-truth data to a new domain without ground-truth.

A. Item identification & Segmentation

Due to aforementioned domain gaps, we trained dedicated segmentation and spatial relationship estimation models on this domain using the same network architecture as in Sec. VI-B1. In contrast, the item identification (ID) model demonstrated comparable performance on VBI images without dedicated training on this domain. This is expected as the training data for the ID model includes reference images obtained from VBI. Therefore, the ID model is trained to encode features from VBI images. As a result, the new segmentation model trained on VBI domain and the same ID model used in the extraction process confirm if each item in a bin is identifiable (`EC1`), and if each item in the bin has at least one item that is not obstructed by other items (`EC2`).

B. Monocular depth estimation

To address the missing depth modality in VBI data, we trained a depth prediction model by fine-tuning a foundational depth prediction network, DepthAnything-v2 [35], [36]. We collected paired RGB-D data of bins with and without bands. When training the depth prediction model, we provide RGB images with bands and provide ground-truth depth maps without bands. We collected synthetic images using Nivida Omniverse [37] to build clean RGB-depth map pairs following approaches from DepthAnything-v2. To build high-quality ground-truth depth maps from real data, we used domain adaption approaches to convert high quality RGB-D data to images captured from VBI towers. We use UNSB [38] that shares core ideas of CycleGAN [39] and CyCada [40] where they use unpaired images from two domains to match styles from one domain to another interchangeably while maintaining contextual features such as boundaries or lines. As shown in Fig. 9, the trained model adds bands realistically, and we can produce RGB images in the target domain. The monocular depth estimation enables us to perform geometrical analysis to understand available free space to avoid the bin lip and metal bar and estimate shapes and orientations of object surfaces.

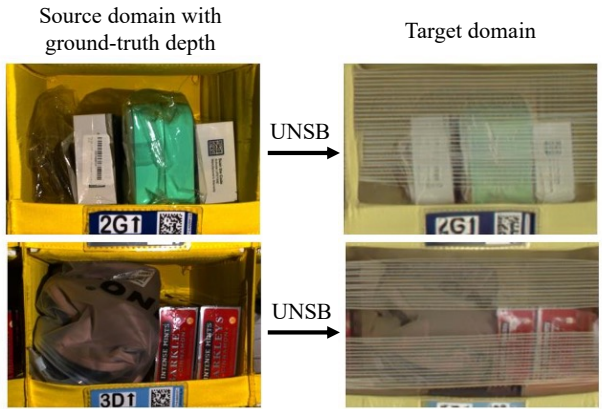


Fig. 9. Examples of results of the domain transfer model converts images captured without bands using our pick camera to the ones captures from VBI towers. The model successfully add artificial bands to the images in the target domain.

C. Movable space estimation

Depth prediction unlocks evaluation of free space around each item to confirm if each item can be lifted to overcome bin lips or moved to the center to avoid metal bars. We compute the free space available within the 3D bin to estimate gaps between items and the bin structures (such as the ceiling of the bin and the metal bar). If the space above items are too narrow to lift the item to overcome bin lips, the items are not eligible to pick. We detect free space along the horizontal direction to check if items behind metal bars can be successfully moved to the center to avoid metal bars. If the items have enough movable space to overcome bin structures, they satisfy this criteria (`EC3`).

D. Pickability prediction

Although the monocular depth prediction network predicts high quality depth maps, the precision and resolutions are not as detailed as the one we use for the core perception pipeline at the station. This prevents us from utilizing the same pick pose generation algorithm using detailed meshes of items. However, visualizations of predicted depth map look close to the one from *Pick camera* with bands not presented. Thus, we use depth maps as a common representation of both domains where we transfer pick poses generated from RGB-D data captured from *Pick camera* to VBI images. To predict pickability scores of objects from depth maps, we design a UNet architecture that takes depth maps as input and predicts pixel-wise probability of pickable surfaces (EC4). We provide segmentation boundaries as additional input to the network to inform boundaries between objects.

As a result, from a monocular RGB image, we apply four eligibility criteria (EC1-4) to filter out picks where items are (1) not identifiable, (2) obstructed, (3) not movable to exit bins, and (4) not pickable by suction. This filtering mechanism enables us to prioritize target items that can be safely handled by our system with confidence. As capacity and reliability of the system grows, we plan to incrementally remove and relax eligibility filters to accept more challenging cases with increased success rates.

IX. Deployment Results

We deployed our system in an active warehouse for more than six months beginning in October 2024. The initial deployment consisted of an extraction system paired with a sortation system in a single station. In February 2025, we updated the configuration by adding a second extraction system to the same station, as shown in Fig. 2. Notably, while maintaining the same footprint as a manual picking station, the dual extraction system configuration effectively doubled the station's picking capacity.

The first phase of deployment primarily focused on integrating the system within existing software architectures in the active warehouse. The integration enabled automated processes for pick request reception, eligibility checking, and data acquisition—including item reference images for identification and dimensional data for bins and items. The system informs pick request successes or failures back to the warehouse software so that new pick requests can be assigned to manual stations when failures happen, which minimizes disruptions to the warehouse operations.

Throughout the deployment period (ending March 2025), the system operated approximately six hours per day on weekdays, and processed more than 12,000 pick requests assigned to the robotic station. Early deployment challenges included various software and hardware issues, such as calibration errors, software communication disruptions, and suction cup damage. These issues were gradually addressed over time, leading to improved system availability.

Once the integration stabilized, we improved core algorithms to fix common failures and minor software bugs to enhance the performance of the system. Following the feature updates in December 2024, we manually annotated each pick attempt to obtain detailed failure analysis discussed in this section. Our performance analysis focuses on the extraction system's capabilities and shows how the system performs the target picking task with success rates of 91%. As the sortation subsystem presented relatively lower operational risks during item conveyance, results discussed in this section primarily focus on performance of the extraction system.

TABLE I
High-level statistics of pick requests

	Number	Rate
Assigned pick requests	6,561	100%
System failure	41	0.6%
Band Separation attempted	6,496	99.0%
Band success	6,403	†98.6%
Band failure	93	†1.4%
Band planning failure (no attempt)	24	0.4%
Item extraction attempted	5,157	78.6%
Extraction success	4,690	*90.9%
Extraction failure	467	*9.1%
Pick planning failure (no attempt)	1,246	19.0%

†, * Rates calculated as proportions of band and pick attempts, respectively.

A. Performance statistics

Table I reports high-level statistics of pick requests from January to March 2025. The system rejected 19.4% of pick requests at the station due to band and pick planning failures and sent them to manual stations. The planning failure happens when the system fails to detect target items with high confidence or fails to find reliable pick poses to extract items with low defect risks. Ideally, those pick requests could have been detected by eligibility check so that the pick requests are not assigned to the station. We discuss primary reasons for rejections in Sec. IX-D.

The system demonstrated 91% pick success when the system attempted to extract target items. This means the system is able to detect correct items from a bin, extract only the target item from the bin, and drop the item safely into the lifting conveyor to deliver the item to totes through the sortation system without defects.

B. Failure cases

The failures and defects include *missed pick* where the system fails to extract the target item from the bin, *multi pick* where the system extracts more than one item, *wrong item* where the system extracts an incorrect item, *drop outside* where the system drops items outside the lifting conveyor, and *damage* if the system damages items.

Figure 10 shows breakdowns of failure and defect cases and a few examples of defect cases observed during the deployment.

Missed pick represents a dominant portion of failures. The primary reasons are weak suction engagement on target surfaces and suboptimal extraction trajectories leading to collision with obstacles like metal bars, bin lips, and bands during the motion. As items are often having their narrow surfaces facing front, it is challenging to engage suction strongly on.

Furthermore, there are items tightly fit to heights of bins. These items are obstructed by bin lips, which exert strong forces that push the items back into the bins.

Multi pick is another large portion of failure cases, and one of the unique challenges faced by extracting items horizontally. *Multi pick* happens when the robot pulls target items and the target items drag adjacent items with friction, or items in unstable conditions fall outside bins when small vibration are applied to them during robot motions.

Wrong item happens relatively less frequently as target items assigned to the station are filtered by the eligibility check pipeline which confirms that items are identifiable from the bin. However, there are items having the same package with minor feature variations, e.g., black ear buds vs. white ear buds, in the same bin. The system mismatches the correct items in such extreme cases. Furthermore, the robot fails to engage suction on a correct item but one of the neighboring items leading to extracting a wrong item.

Damage cases are rare for our system as the system does not pull items against gravity. As such, sliding items through the surfaces of bins is less likely to open lids of packages or deconstruct items into multiple pieces. However, when the suction cup holds items at a higher location above the lifting conveyor, lids of boxes are opened as the weight of items are not supported. Additionally, the *band robot* has to pull elastic bands with high force. When an item is located in the middle of the band separation trajectory, the item is often crushed by the hook on the *band EoAT*, which is one of the primary root causes of severe damage.

Drop outside Items can bounce off the conveyor when dropped from excessive heights, especially when combined with rapid pulling motions. While the conveyor height is typically set high enough to safely receive extracted items, the current height setting is not optimal to reduce risks of items bouncing. The height of the lifting conveyor and item drop motion will be further optimized to minimize these risks.

C. Impact of continuous visual feedback

The continuous tracking of item locations enables the robot to adapt motions as soon as failure is detected. As reported in Table II, this is one of the key features that lifted the success rate of the system by 5%. Thanks to item tracking, it is possible to plan a new trajectory to locate the EoAT on the target items even though their locations or orientations are significantly changed from their initial states.

TABLE II
SUCCESS RATE WITH AND WITHOUT ONLINE MOTION ADAPTATION WITH VISUAL FEEDBACK

	Attempts	Success	Success Rate
No online adaptation	5,157	4,447	86%
With online adaptation	5,157	4,690	**91%**

D. Eligibility check and Pick planning failure

For 19% of all pick requests, the system fails to build a plan to pick a target item after successfully opening bands for a target bin. The breakdowns of the reasons are the following (percentages are over all pick requests):

- No ID or Low ID confidence (3%): Item identification is required to proceed to the next planning steps, and the system often fails to detect target items, or produces low confidence scores if the feature embedding is not close to the target item. There are also cases where target item is not visible due to occlusion.
- No grasp pose without collision (8%): The pose generator often fails to find pick poses having low risk of failures without collision. It happens frequently on items with narrow surfaces facing front. There are also cases where pick poses are generated but the system fails to find a collision-free trajectory for pick pose candidates if the pick pose is close to the band robot or lifting conveyor structures.
- No item having good spatial status (7%): Based on the spatial status of items predicted from the core perception pipeline, the system avoids picking stacked items below others, or blocked items behind others. If the target item is detected as stacked or blocked, the system does not generate pick poses.

The reasons above are strongly correlated with the eligibility criteria discussed in Sec. VIII as eligibility check pipeline checks if items are (1) identifiable, (2) pickable, and (3) obstructed. When eligibility check was not deployed, the portion of pick planning failure was more than 60%, and accepting only eligible pick requests has successfully reduced the rate to lower than 20%.

When a planning failure happens, the extraction cycle is immediately cancelled to release the pod. The failed pick request is re-assigned to manual stations. As relocating pods from robotic stations to manual stations via mobile robots could introduce congestion to the floor, we need to reduce the number of planning failures with improved eligibility filters.

X. LESSONS LEARNED & INSIGHTS

The extensive deployment of our system has provided valuable insights into both the practical challenges and effective solutions. Through this real-world test, we have identified solutions work well in the real world and approaches that performed worse than expected. We share these findings with the broader robotics community to highlight future directions for developing more reliable and scalable automated systems.

1) Continuous observation through an eye-in-hand camera and item tracking: The role of continuous feedback has been overlooked in robotic manipulation applications such as bin picking or package picking. In classical bin picking tasks, the interaction between robots and objects does not change the surrounding configuration significantly as robots are supposed to pick objects not obstructed by other objects, objects are mostly rigid, and object shapes are known from CAD models. Thus, open-loop control based on initial perception results has been sufficient. However, picking targeted objects, previously unseen objects, from densely packed containers increase uncertainties significantly as soon as a robot starts touching objects. Continuous visual feedback successfully handles such

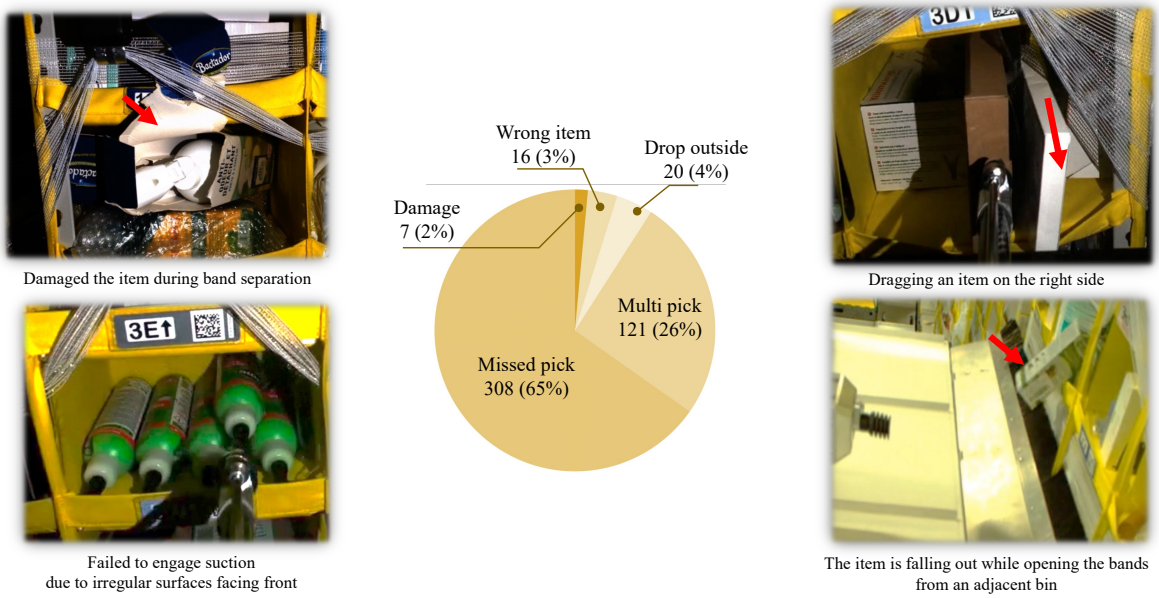


Fig. 10. Breakdowns of failure modes and examples of a damage case (top-left), a missed pick (bottom-left), and multi pick cases (right).

uncertainties by fixing errors from initial perception results or by adapting behavior as objects move during robot-item interactions. This contributes to additional reliability of the system leading to higher success rates. Recent achievements in video object tracking models [33], [41] enable use of accurate object tracking results that are robust to occlusions in real-time. Thus, we recommend utilizing robust tracking results to tackle complex manipulation tasks.

2) SDF as core representation for manipulation: Our system extensively uses the SDF representation of the scene for multiple purposes such as 3D mesh reconstruction, fast collision checking during pick pose generation, movable space estimation within a bin, quality estimation of suction poses, and detecting the risk of touching adjacent items. The SDF with projected object labels represents 3D space in a queryable format such that it can check if a point (x, y, z) is touching an object or not by reading the value in the 3D array. In comparison to classic tree search, nearest neighbor search algorithms, SDF is much faster and more informed as SDF values represent distances to the closest surfaces and directions. One of the insights we exploit is that the 3D region of interest can be bounded by the maximum size of the bins. This enables us to use a voxel size of $4\text{x}4\text{x}4$ mm to represent fine-grained details of 3D scenes in a 128x128x128 grid. GPU-based parallelism further enhanced the utilization of SDF when building and processing SDF values. The SDF-based representation also unlocks fusing data from multiple cameras or multiple view points, and we observed enhanced coverage of object surfaces when taking multiple views. We believe SDF is an efficient way to represent the 3D world for robotic manipulation tasks.

3) Spatial reasoning between objects via monocular image: We initially used 3D pointclouds of individual items to estimate spatial relationships between objects such as if an item is below others or behind others. However, we found that it is important to understand which boundary belongs to which item to confirm if the item is obstructed or not. The obstructed items have boundaries owned by other items while unobstructed items have fully visible boundaries. This is difficult to derive from a 3D pointcloud without color and texture information for each item. This motivated us to train a model using monocular RGB images to classify spatial status of individual items. As a result, the segmentation model with a spatial estimation head introduced in Sec. VI-B1 successfully distinguished confusing cases leading to improved classification performance (from 86%/87% to 97%/96% in terms of precision and recall).

4) Prevention of items falling out in shelf-style container picking: The high frequency of *Multi pick* cases highlights a critical challenge in shelf-style container manipulation: items frequently fall out during operations. This is particularly problematic when items are in unstable poses, as they can dislodge either during band opening or from minor disturbances to the pods. While our lifting conveyor successfully prevents severe damage by catching fallen items, returning these items to their bins introduces additional operational costs. This observation underscores the need for more sophisticated preventive mechanisms to address *Multi pick* scenarios before they occur.

5) Holistic approaches in robotic system design: The design choices presented in this paper resulted from comprehensive discussions involving experts across various system components, including hardware, software, perception, and motion planning & control. For instance, during early exploration, we encountered challenges with items falling to the floor due to high momentum upon exiting the bins. While this could have been addressed by adding extra modalities to the *pick EoAT* (such as a fingered gripper) or through motion optimization, we instead developed the lifting conveyor solution. This approach not only simplified the problem but also yielded additional benefits, including reduced cycle times.

By allowing the extraction system to prepare for the next bin immediately after dropping items onto the lifting conveyor, the *pick robot* no longer needed to move to a specific drop-off position after each pick. This systematic solution significantly mitigated damage risks and reduced the number of items falling to the floor.

XI. Outlook

We outline our planned next steps to address the identified challenges for continued improvement of robotic picking systems.

A. Opportunities for learned visuomotor policies

The perception and motion approaches in Section VI are the result of an iterative design loop in which we improved our system by adding or refining behavior to address the current most dominant failure case. This strategy has proven fruitful in our early development stages when we encountered clusters of frequently occurring failure cases. As error root causes grow increasingly specialized and require more tailored solution strategies, we expected this design approach to become less efficient.

This motivated us to explore recent advances in visuomotor policy learning (VMP), such as diffusion policies [42] and flow matching [43], as an alternative to the current hand-crafted robot behaviors.

Our initial explorations using simulated training data confirmed that VMPs can deliver comparable success rates to our hand-crafted motion strategies on real robot tests. Additionally, VMPs demonstrated adaptive behavior to recover from near-failure situations on challenging cases such as picking items from the bottom of a vertical item stack.

A key challenge of deploying learned VMPs is their lack of interpretability in failure cases. While it is often easy to fix a simple bug or improve algorithmic limitations of heuristic approaches, VMPs have to be re-trained or fine-tuned to learn about failures while maintaining their previous performance. Preliminary explorations showed that it is possible to fine-tune VMPs on specific failure cases by reproducing real-world failures in simulation using a Real2Sim module. Therefore, we are continuing to explore a scalable Real2Sim pipeline to enable autonomous VMP improvement from failures. We anticipate that further investigations in continual learning of robot behavior through failures can play an essential role in solving long tails of failure cases at scale.

B. Additional modality to the pick EoAT

The current *pick EoAT* has a limitation when it has to pick items having narrow surfaces only accessible from the front. We explored a prototype of a blade having vacuum channels on its surface to pull items via side-facing surfaces of objects (see Fig. 11). The new modality requires more advanced motion and control to insert the blade between items. We plan to integrate the new modality to the *pick EoAT* in the future. We expect to reduce missed pick cases and planning failure cases by using the new modality.

Fig. 11. Side-way suction on a blade-style tool. We explored pick poses and motion strategies to utilize the new modality to pick narrow items using their side surfaces.

XII. Conclusion

This paper presented an end-to-end robotic system designed for autonomous picking of targeted objects from fabric pods in production warehouse environments. Our solution combines proven classical methods with state-of-the-art approaches in computer vision and robot motion control, along with specialized hardware optimized for the task. Through six months of deployment processing over 12,000 customer orders, we demonstrated that the system can achieve more than 90% success rate.

The key contributions of our work include: (1) perception and control algorithms that enable both spatial understanding of item relationships and continuous tracking for failure recovery through visual feedback, (2) the eligibility concept that identifies target items that are eligible to the robotic system with domain adaptation techniques, (3) specialized EoATs optimized for separating bands and extracting items from shelf-style containers and transporting items from the conveyor, and (4) comprehensive deployment insights from processing over 12,000 customer orders, including detailed performance analysis, common failure modes, and lessons learned for scaling robotic picking systems.

Our deployment experience revealed several important insights for developing reliable robotic systems at scale. We believe our work demonstrates the feasibility of robotic picking operations in production environments while highlighting important challenges and opportunities for future research in this domain. By sharing our experiences and insights from a large-scale, real-world deployment, we aim to encourage collaborative efforts that push the boundaries of robotic manipulation in industrial settings.

References

[1] C. Eppner, S. Höfer, R. Jonschkowski, R. Martín-Martín, A. Sieverling, V. Wall, and O. Brock, "Lessons from the amazon picking challenge: Four aspects of building robotic systems." in *Robotics: Science and Systems*, vol. 12, 2016.

[2] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman, "Analysis and observations from the first amazon picking challenge," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 1, pp. 172–188, 2016.

[3] D. Morrison, A. W. Tow, M. Mctaggart, R. Smith, N. Kelly-Boxall, S. Wade-Mccue, J. Erskine, R. Grinover, A. Gurman, T. Hunn *et al.*, "Cartman: The low-cost cartesian manipulator that won the amazon robotics challenge," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 7757–7764.

[4] R. Memmesheimer, J. Nogga, B. Pätzold, E. Kruzhkov, S. Bultmann, M. Schreiber, J. Bode, B. Karacora, J. Park, A. Savinykh *et al.*, "Robocup@ home 2024 opl winner nimbro: Anthropomorphic service robots using foundation models for perception and planning," *arXiv preprint arXiv:2412.14989*, 2024.

[5] C. G. Atkeson, P. B. Benzun, N. Banerjee, D. Berenson, C. P. Bove, X. Cui, M. DeDonato, R. Du, S. Feng, P. Franklin *et al.*, "What happened at the darpa robotics challenge finals," *The DARPA robotics challenge finals: Humanoid robots to the rescue*, pp. 667–684, 2018.

[6] E. Krotkov, D. Hackett, L. Jackel, M. Perschbacher, J. Pippine, J. Strauss, G. Pratt, and C. Orlowski, "The darpa robotics challenge finals: Results and perspectives," *The DARPA robotics challenge finals: Humanoid robots to the rescue*, pp. 1–26, 2018.

[7] N. Hudson, F. Talbot, M. Cox, J. Williams, T. Hines, A. Pitt, B. Wood, D. Frousheger, K. L. Surdo, T. Molnar *et al.*, "Heterogeneous ground and air platforms, homogeneous sensing: Team csiro data61's approach to the darpa subterranean challenge," *Field Robotics*, vol. 2, pp. 595–636, 2022.

[8] K. Ebadi, L. Bernreiter, H. Biggie, G. Catt, Y. Chang, A. Chatterjee, C. E. Denniston, S.-P. Deschênes, K. Harlow, S. Khattak *et al.*, "Present and future of slam in extreme environments: The darpa subt challenge," *IEEE Transactions on Robotics*, vol. 40, pp. 936–959, 2023.

[9] Amazon Robotics, "Amazon's robot arms break ground in safety and technology," 2021, accessed: 2025-03-01. [Online]. Available: https://www.amazon.science/latest-news/amazon-robotics-see-robin-robot-arms-in-action

[10] About Amazon, "Amazon introduces sparrow — a state-of-the-art robot that handles millions of diverse products," 2022.

[11] "Covariant robotic induction," https://covariant.ai/robotic-induction [Accessed: 09/04/2025].

[12] "Berkshire grey robotic picking," https://www.berkshiregrey.com/solutions/robotic-pick/ [Accessed: 09/04/2025].

[13] "Introducing next-generation rightpick 3 item-handling robot system," https://righthandrobotics.com/the-latest/introducing-next-generation-rightpick-3-item-handling-robot-system [Accessed: 09/04/2025].

[14] "Xyz robotics piece picking robot," https://www.xyzrobotics.com/picking-robot/piece-picking [Accessed: 09/04/2025].

[15] M. Bajracharya, J. Borders, R. Cheng, D. Helmick, L. Kaul, D. Kruse, J. Leichty, J. Ma, C. Matl, F. Michel *et al.*, "Demonstrating mobile manipulation in the wild: A metrics-driven approach," in *Robotics: Science and Systems*, 2023.

[16] E. Goldman, R. Herzig, A. Eisenschtat, J. Goldberger, and T. Hassner, "Precise detection in densely packed scenes," in *Proc. Conf. Comput. Vision Pattern Recognition (CVPR)*, 2019.

[17] K. Thomas Wilson, "Automated guided vehicles for material flow in fulfillment centers," Ph.D. dissertation, Massachusetts Institute of Technology, 2023.

[18] Amazon Fulfillment Technologies, "Amazon bin image dataset," 2018. [Online]. Available: https://registry.opendata.aws/amazon-bin-imagery/

[19] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 3041–3050.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[22] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *NeurIPS*, 2021.

[23] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022.

[24] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE international symposium on mixed and augmented reality*. Ieee, 2011, pp. 127–136.

[25] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *SIGGRAPH Comput. Graph.*, vol. 21, no. 4, p. 163–169, Aug. 1987. [Online]. Available: https://doi.org/10.1145/37402.37422

[26] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[27] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[28] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[29] A. Jagannathan and E. L. Miller, "Three-dimensional surface mesh segmentation using curvedness-based region growing approach," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 29, no. 12, pp. 2195–2204, 2007.

[30] J. Papon, A. Abramov, M. Schoeler, and F. Worgotter, "Voxel cloud connectivity segmentation-supervoxels for point clouds," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2027–2034.

[31] S. Christoph Stein, M. Schoeler, J. Papon, and F. Worgotter, "Object partitioning using local convexity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 304–311.

[32] S. Li, A. Keipour, K. Jamieson, N. Hudson, C. Swan, and K. Bekris, "Large-scale package manipulation via learned metrics of pick success," in *Robotics science and systems*. Robotics: Science and Systems, 2023.

[33] H. K. Cheng, S. W. Oh, B. Price, J.-Y. Lee, and A. Schwing, "Putting the object back into video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3151–3161.

[34] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.

[35] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21 875–21 911, 2025.

[36] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 371–10 381.

[37] NVIDIA Corporation, *NVIDIA Omniverse Documentation*, 2025. [Online]. Available: https://docs.nvidia.com/omniverse/index.html

[38] B. Kim, G. Kwon, K. Kim, and J. C. Ye, "Unpaired image-to-image translation via neural schrödinger bridge," in *The Twelfth International Conference on Learning Representations*, 2024.

[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[40] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. Pmlr, 2018, pp. 1989–1998.

[41] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[42] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[43] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.