**Advancing Robotic Vision for Warehouse Bin-Picking: A Thesis Proposal Guide**

**Executive Summary**

The burgeoning demands of e-commerce and the inherent complexities of manual labor have underscored the critical need for advanced automation in warehouse logistics. Robotic bin-picking, a cornerstone of this automation, presents significant challenges, particularly in environments characterized by dense clutter, severe occlusion, and a diverse array of known items. This report provides a comprehensive review of the current state of research in robotic perception for bin-picking, focusing on the utilization of RGB and RGB-D cameras from a top-down perspective, techniques for few-shot and one-shot 3D shape learning, and an overview of relevant datasets and simulation environments.

Current perception modules effectively fuse RGB-D inputs to create structured scene representations, with approaches like U-Net architectures demonstrating high precision in predicting grasp regions directly from RGB-Points data, bypassing explicit pose estimation. However, the fixed top-down camera setup, common in industrial applications, necessitates robust inference mechanisms to overcome severe occlusions without active viewpoint adjustment. The report highlights the emerging role of data-efficient learning paradigms, such as object-centric adaptation (e.g., ControlVLA) and zero-shot generalization via Vision-Language Models (e.g., ObjectVLA), which enable robots to learn from minimal demonstrations and generalize to novel objects. Furthermore, the two-stage learning approach of AnyDexGrasp exemplifies how contact-centric representations can lead to human-level learning efficiency in dexterous grasping.

Existing industrial datasets, such as XYZ-IBD, reveal a substantial "reality gap" where state-of-the-art methods struggle with the unique challenges of texture-less, metallic, and symmetrical objects in dense clutter. This emphasizes the critical value of real-world data collection, complemented by sophisticated synthetic data generation frameworks like RoboTwin 2.0, which employ extensive domain randomization and sim-to-real transfer techniques (e.g., RCANs, GANs) to bridge this gap.

Leveraging the unique context of a planned real-world dataset collection (top-down PhotoNeo RGB-D camera, known items, cross-bin inference), this report proposes three novel and under-explored MSc thesis directions. These directions aim to advance the field by focusing on probabilistic 3D shape reconstruction from disparate partial views, few-shot grasp success prediction for occluded industrial items, and learning robust semantic-geometric features under extreme occlusion. These research avenues offer significant contributions to developing more adaptable and efficient robotic systems for industrial logistics.

**1. Introduction to Robotic Bin-Picking for Warehouse Logistics**

**1.1. The Imperative of Automation in Modern Warehouses**

The landscape of modern logistics is undergoing a profound transformation, largely driven by the exponential growth of e-commerce and persistent labor shortages. This shift places an unprecedented demand on warehouses to process an ever-increasing volume of goods with greater speed and efficiency. Within this context, automation is no longer merely an advantage but a fundamental necessity. Robotic systems are increasingly deployed to handle repetitive, physically demanding tasks, thereby alleviating human workers from tedious workloads and enhancing overall operational throughput.1

Robotic bin-picking stands out as a particularly critical area of automation. It addresses the bottleneck of manually retrieving individual items from bulk containers, a process that is both labor-intensive and prone to error. In high-throughput environments, such as a production warehouse processing approximately 1200 bins per day, automating this task is paramount for maintaining operational fluidity and scalability. The ability of robots to accurately and reliably pick diverse items from bins directly impacts the efficiency of downstream processes like sorting, packing, and order fulfillment.

**1.2. Defining the Bin-Picking Challenge: Clutter, Occlusion, and Item Variability**

Despite its immense potential, robotic bin-picking remains a significant challenge due to the inherent complexities of unstructured environments. Objects within a bin are typically randomly oriented, often overlapping, and severely occluded by one another or the bin's structure.2 This disordered arrangement makes precise object recognition, pose estimation, and grasp planning exceptionally difficult for autonomous manipulators.2

The challenge is further compounded by the diversity of items encountered in a typical warehouse setting. The scenario involves up to five known items, typically from distinct object categories such as t-shirts, bottles, and boxes. This variability in shape, size, material properties, and deformability adds layers of complexity to the perception and manipulation tasks compared to more homogeneous industrial parts.2 While knowing the specific items in the catalogue simplifies the problem by eliminating the need for open-set detection, the distinct categories imply significant inter-category differences in visual and physical properties. This creates a unique problem where solutions must be robust to the wide range of appearances and geometries across categories, while simultaneously leveraging any intra-category consistency (e.g., a canonical representation for each known item type). This problem formulation suggests that successful approaches will need to learn robust, generalized representations for each known category, moving beyond purely instance-specific models. The ability to handle such diverse objects within a cluttered environment is crucial for real-world deployment in logistics automation.2

**2. Current State of Robotic Perception in Bin-Picking**

**2.1. RGB and RGB-D Camera Utilization from a Top-Down Perspective**

The choice of sensing modality is fundamental to effective robotic perception in bin-picking. RGB-D cameras, which provide both color (RGB) and depth information, have become indispensable tools for achieving robust scene understanding in complex, three-dimensional environments.

**2.1.1. Advantages of RGB-D for Bin-Picking Perception**

RGB-D cameras, such as the PhotoNeo PhoXi or MotionCam-3D, offer a significant advantage over traditional RGB cameras by providing crucial geometric information alongside visual appearance cues. The RGB component delivers details about surface properties and material types, which are vital for distinguishing objects and identifying features.5 Simultaneously, the depth channel captures spatial geometry, enabling the system to understand occlusion, volume, and the spatial relationships between objects.5 This dual input is particularly beneficial in cluttered bin-picking scenarios where objects are often stacked, occluded, or have low texture, making traditional 2D vision insufficient for accurate 3D pose estimation and grasp planning. The depth data directly assists in segmenting individual objects, estimating their three-dimensional positions and orientations, and identifying viable surfaces for grasping, even under challenging lighting conditions.5

**2.1.2. Techniques for Handling Partial Occlusion and Overlapping Items**

Addressing partial occlusion and overlapping items from a top-down perspective is a central challenge in bin-picking. Raw depth data can often be suboptimal for object-level reasoning due to variations across different surfaces of the same object.5 To overcome this, advanced perception pipelines incorporate modules like Object Depth Mapping (ODM), which fuse raw depth images with object masks derived from RGB images.5 Instead of relying on pixel-wise depth, ODM computes the average depth for each object, generating a simplified, object-centric spatial representation that is more conducive to downstream decision-making.5

Another effective strategy involves directly predicting graspable regions or suction points from fused RGB-D data, thereby bypassing the often-difficult and computationally intensive steps of full object recognition and 6D pose estimation in heavily cluttered scenes.2 U-Net based Convolutional Neural Networks (CNNs) have shown remarkable efficacy in this regard. These networks combine RGB images and depth information (or point clouds derived from depth) to output a probability map indicating the likelihood of successful grasping at each pixel.2 Notably, using RGB-Points as input, which converts depth information into a dense point cloud and fuses it with RGB, has demonstrated superior precision, achieving up to 95.74% accuracy in predicting picking regions.2 This observation suggests that for a PhotoNeo RGB-D camera, converting depth maps into point clouds for perception processing should be a primary consideration, as it offers a demonstrably higher precision for grasp region prediction, highlighting the importance of point cloud processing architectures.

While next-best-view (NBV) strategies can significantly improve perception in occluded scenes by allowing the robot to actively select more informative viewpoints 5, a fixed top-down camera setup, as planned, precludes such active exploration. This fixed viewpoint presents a critical constraint: the perception system must be exceptionally robust to severe occlusions and partial observability from a single, static perspective. The challenge shifts from active information gathering to passive, intelligent inference from limited data. This points to a significant research opportunity: developing algorithms that can infer complete object information from incomplete visual cues, such as implicit shape representations or physics-informed reconstruction, without the benefit of dynamic camera movement.

**Table 1: Key Perception Techniques for Top-Down Bin-Picking**

| Technique | Input Modality | Occlusion Handling | Output | Key Strengths | Relevance to User's Setup |
| --- | --- | --- | --- | --- | --- |
| U-Net (Region Prediction) | RGB-D, RGB-Points | Direct prediction of graspable regions | Probability Map for Grasping Region | Data efficient, robust to clutter, high precision with RGB-Points | High precision with RGB-Points, suitable for object-agnostic grasping |
| XPG-RL (Perception Module) | Fused RGB-D | Object Depth Mapping (ODM) for simplified spatial representation | Structured Scene Representation | Improved efficiency in cluttered scenes, object-centric approach | Object-centric approach for fixed top-down view |
| GeoMVSNet (3D Reconstruction) | Multi-view RGB | Integrates geometric priors from coarse depth maps | Dense Depth Maps/Point Clouds | Precise depth estimation, robust aggregation in finer stages | Potential for robust 3D reconstruction from multiple partial views (cross-bin) |
| One-Shot Neural Fields (NeRFs) | Single/Few RGB | Recovers occluded object parts, novel view rendering | Latent Code for Geometry/Appearance, Grasp Prediction | Compact representation, 3D reconstruction from limited views, stable grasp prediction | Potential for robust 3D understanding from fixed top-down view, grasp prediction |

**2.2. Techniques for Few-Shot or One-Shot 3D Shape Learning and Grasp Prediction**

The efficiency of data acquisition in real-world robotic manipulation remains a significant bottleneck. Consequently, few-shot or one-shot learning paradigms, which enable systems to generalize from minimal demonstrations, are gaining prominence.

**2.2.1. Generalization from Limited Demonstrations**

A core challenge in robotic manipulation is the ability to learn and generalize tasks when only a limited number of demonstrations are available.7 Traditional methods often require extensive datasets, sometimes involving millions of grasp trials, to achieve satisfactory performance.2 This data scarcity in real-world settings necessitates approaches that can efficiently transfer knowledge and adapt to novel objects and scenarios with minimal new data.

**2.2.2. Object-Centric Adaptation for Efficient Fine-Tuning (e.g., ControlVLA)**

Frameworks like ControlVLA address the challenge of data scarcity by bridging pre-trained Vision-Language-Action (VLA) models with object-centric representations.7 These models leverage the rich prior knowledge embedded in large-scale VLA pre-training and adapt general-purpose manipulation policies to specific tasks with remarkable data efficiency, requiring as few as 10-20 demonstrations to achieve high success rates (e.g., 76.7% across diverse tasks).7 Object-centric representations reduce the complexity of the input observation space by focusing on relevant object properties (shape, size, position), enhancing policy robustness to changes in object pose and instance and making them less susceptible to real-world noise.7 This approach is particularly relevant for fine-tuning models on specific known items, as it maximizes the impact of a limited but valuable real-world dataset.

**2.2.3. Zero-Shot Object Generalization via Vision-Language Models (e.g., ObjectVLA)**

ObjectVLA represents a promising approach for enabling robots to generalize learned skills to novel objects without requiring explicit human demonstrations for each new target.10 This framework curates visual-textual pairs augmented with localization metadata (e.g., bounding boxes) and co-finetunes this data with teleoperated robot interaction data. By embedding localization as a bridging representation, ObjectVLA creates a unified pathway between visual-language inputs and robotic actions, enabling zero-shot object generalization.10 The model has demonstrated a 64% success rate on 100 novel objects not seen during training, and can even adapt rapidly to new objects by collecting smartphone-captured images for lightweight fine-tuning.10 This capability is highly beneficial for semantic understanding and recognition of known items, even with partial views.

**2.2.4. Two-Stage Learning for Dexterous Grasping (e.g., AnyDexGrasp)**

AnyDexGrasp introduces a two-stage learning methodology for dexterous grasping that achieves human-level learning efficiency.13 The first stage involves training a universal, hand-agnostic model to map scene geometry to intermediate contact-centric grasp representations (CGRs). This stage requires a large-scale dataset but is performed offline and is independent of specific robotic hands. The second stage then trains a unique grasp decision model for each robotic hand, translating these CGRs into final grasp poses through real-world trial and error.13 This second stage is remarkably data-efficient, requiring only hundreds of grasp attempts on a modest set of training objects (e.g., 40 objects).13 The approach demonstrates robust generalization, achieving 80-98% grasp success on over 150 previously unseen objects in cluttered environments.13 This efficiency stems from a geometry coverage analysis that shows scaling up the sample density of local geometries per object is more impactful than simply increasing the number of training objects, allowing a wide range of local geometries to be covered with fewer unique items.13

**2.2.5. One-Shot 3D Object Understanding with Neural Fields (e.g., NeRFs)**

Neural Radiance Fields (NeRFs) offer a powerful approach for 3D object understanding by representing a scene as a continuous volumetric function that assigns color and volume density to any point in space.17 NeRFs can learn category-level priors from large multi-view datasets and then fine-tune on novel objects from a single or few views at test time.18 This compact scene representation allows for various tasks, including novel view rendering, 3D reconstruction (e.g., recovering depth or point clouds), collision checking, and stable grasp prediction.18 Critically, NeRFs can recover representations that allow rendering from novel views, including occluded object parts, and directly decode grasp poses from their latent representation.18 This capability is highly relevant for inferring complete 3D information from limited, fixed-view observations in cluttered bins.

The user's plan to collect a month's worth of real-world data (potentially hundreds of thousands of object instances across ~36,000 bins) is substantial, yet it does not equate to the "millions of grasp trials" often required by traditional methods.2 The few-shot and one-shot learning methods described (ControlVLA, ObjectVLA, AnyDexGrasp) are specifically designed for such data-efficient scenarios. This suggests that the collected data is perfectly suited for fine-tuning pre-trained models or for training the more data-efficient stages of two-stage methods. This approach maximizes the impact of the collected dataset.

Given the fixed top-down camera and the presence of partial occlusion and overlapping items, the system will inherently receive fragmented visual information. NeRFs can reconstruct 3D from single or few RGB inputs 18, and AnyDexGrasp uses contact-centric representations derived from partial observations.13 This suggests that instead of attempting to reconstruct a full CAD model, the system could learn an implicit 3D shape representation for each known item from various partial views observed across different bins. This implicit representation could then be effectively used for matching to a known catalogue and predicting grasp success. This approach sidesteps the difficulty of explicit 3D reconstruction from heavily occluded, single-view data, offering a more robust path to 3D understanding for grasp prediction in challenging bin-picking scenarios.

**Table 2: Few-Shot/One-Shot 3D Learning and Grasp Prediction Methods**

| Method | Core Idea | 3D Shape Learning | Grasp Prediction | Data Efficiency | Generalization | Relevance to User's Scenario |
| --- | --- | --- | --- | --- | --- | --- |
| ControlVLA | Object-centric adaptation for efficient fine-tuning of VLA models | Object-centric representations | Policy adaptation for specific tasks | 10-20 demonstrations | Robust to unseen objects/backgrounds | Efficient fine-tuning for known items with limited real data |
| ObjectVLA | VLA for zero-shot object generalization via localization-aware reasoning | Localization-aware reasoning, visual-textual pairs | Action prediction via unified pathway | Zero-shot (via VLM pre-training) | 64% success on 100 novel objects | Leveraging VLM for semantic understanding, adaptable with smartphone images |
| AnyDexGrasp | Two-stage: hand-agnostic representation + hand-dependent grasp decision | Contact-centric grasp representations (CGRs) from partial observations | Grasp decision model (quality estimation) | Hundreds of attempts on 40 objects | 80-98% on 150 unseen objects in clutter | Practical for limited real-world data, robust in cluttered scenes |
| One-Shot Neural Fields (NeRFs) | Implicit 3D scene representation from limited views | Latent codes for geometry and appearance, 3D reconstruction from single/few views | Implicit grasp decoder | Single/few RGB inputs | Robust robotic grasping, novel view rendering of occluded parts | 3D understanding from fixed top-down view, grasp prediction |

**3. Relevant Datasets, Simulation Environments, and Benchmarks**

The development and evaluation of robotic bin-picking systems heavily rely on access to diverse and representative datasets, as well as flexible simulation environments for scalable data generation and testing.

**3.1. Overview of Existing Datasets for Bin-Picking**

**3.1.1. Industrial-Grade Challenges (e.g., XYZ-IBD)**

Existing datasets often fall short in capturing the full spectrum of complexities encountered in real-world industrial bin-picking scenarios. The XYZ-IBD dataset was specifically developed to address this gap, featuring 15 texture-less, metallic, and mostly symmetrical industrial objects that vary in shape and size.19 These objects are randomly and densely arranged in bins, leading to severe occlusion and dense clutter, which are characteristic of authentic industrial conditions.19 The dataset provides millimeter-accurate 6D pose annotations, collected through a meticulous semi-automatic pipeline involving anti-reflection spray, multi-view depth fusion, and ICP refinement.20 Benchmarking state-of-the-art methods on XYZ-IBD has revealed a stark performance drop compared to their performance on more common household object datasets, highlighting a significant gap between current academic benchmarks and the realities of industrial environments.19 This performance degradation underscores that the "reality gap" is particularly pronounced and currently unresolved for industrial bin-picking, especially with objects exhibiting challenging properties like high reflectivity and symmetry. This implies that real-world data collection in a production warehouse is not merely supplementary; it is critical for addressing this specific, challenging domain. Any synthetic data generation must be meticulously designed to mimic these industrial complexities, and sim-to-real methods must be specifically evaluated for their effectiveness on such challenging object properties, pointing to a strong need for research focused on robust perception for these difficult industrial materials.

**3.1.2. Large-Scale Object Libraries (e.g., RoboTwin-OD)**

Beyond specific bin-picking challenges, large-scale object libraries are crucial for training and evaluating generalizable manipulation policies. RoboTwin-OD, part of the RoboTwin 2.0 framework, is an extensive object library comprising 731 instances across 147 categories.23 Each object is richly annotated with semantic and manipulation-relevant labels, including placement points, functional points, grasping points, and grasp axis directions, along with diverse language descriptions.26 This comprehensive library, constructed from various sources and augmented with a library of 12,000 high-quality textures, provides an invaluable resource for creating diverse training scenarios and benchmarking robotic manipulation tasks.26

**3.2. Simulation Environments for Data Generation and Sim-to-Real Transfer**

The high cost and time commitment associated with collecting and labeling real-world robotic data make simulation environments an attractive alternative for generating large-scale datasets.27

**3.2.1. Scalable Data Generation with Domain Randomization (e.g., RoboTwin 2.0)**

RoboTwin 2.0 is a scalable simulation framework designed for automated, large-scale generation of diverse and realistic data for robotic manipulation.23 It incorporates structured domain randomization along five key axes: clutter, lighting, background, tabletop height, and language instructions.23 This systematic randomization enhances data diversity and policy robustness, significantly improving sim-to-real transfer.23 Policies trained exclusively on RoboTwin 2.0's synthetic data have demonstrated strong generalization capabilities to unseen real-world tasks, achieving a 228% relative gain in zero-shot performance.23 This framework provides a robust blueprint for generating synthetic data that can effectively augment real-world observations.

**3.2.2. Bridging the Reality Gap: Advanced Sim-to-Real Techniques**

Despite advancements in synthetic data generation, a persistent "reality gap" exists between simulated and real-world environments, often leading to performance degradation when models trained in simulation are deployed in reality.27 To bridge this gap, advanced sim-to-real techniques are crucial.

Randomized-to-Canonical Adaptation Networks (RCANs) offer a novel approach by learning to translate heavily randomized simulated images into their non-randomized, canonical versions.27 This learned adaptation function then allows real-world images to be transformed into a format that resembles the canonical simulation, enabling policies trained purely in simulation to operate effectively in the real world.27 RCANs demonstrate remarkable data efficiency, achieving 70% zero-shot grasp success on unseen objects and reducing the need for real-world finetuning data by over 99% compared to state-of-the-art systems.27

Generative Adversarial Networks (GANs), specifically CycleGAN, are also employed to improve the photorealism of synthetic images, directly reducing the reality gap.28 This involves training a GAN to translate synthetic images to appear more realistic, and vice-versa, without requiring paired synthetic and real images.28 While effective, challenges include ensuring the translation preserves the semantics crucial for the task (e.g., not erasing objects in shadows) and mitigating image defects like checkerboard artifacts or blurring.28

The user's ability to collect real-world data and generate synthetic data presents a unique opportunity. The challenges of real-world data collection (costly, time-consuming manual labeling) and the limitations of synthetic data (oversimplified environments, reality gap) are well-documented.23 However, advanced sim-to-real methods like RCANs and GANs can significantly bridge this gap. Therefore, the optimal strategy is a hybrid data approach. The real-world data should be collected to capture true industrial complexity and used for validation and targeted fine-tuning. Concurrently, large-scale synthetic data should be generated with aggressive domain randomization, mimicking RoboTwin 2.0's principles, and sophisticated sim-to-real techniques should be employed to ensure models trained on synthetic data generalize effectively to the real world. This approach minimizes the burden of real-world annotation while maximizing data diversity.

**Table 3: Key Datasets and Simulation Environments for Robotic Bin-Picking**

| Name | Type | Object Characteristics | Clutter/Occlusion | Annotation Quality | Key Features | Relevance to User's Project |
| --- | --- | --- | --- | --- | --- | --- |
| XYZ-IBD | Real-world, Hybrid (Real + Synthetic) | 15 texture-less, metallic, symmetrical industrial objects | Severe occlusion, dense clutter, reflections | Millimeter-accurate 6D pose | Industrial complexity, benchmark for 6D pose, multi-view real/synthetic data | Directly addresses industrial challenges, provides benchmark for validation |
| RoboTwin 2.0 (with RoboTwin-OD) | Synthetic | 731 instances, 147 categories, semantic/manipulation labels | Structured domain randomization (clutter, lighting, background, tabletop height, language) | Semantic and manipulation-relevant labels | Scalable data generation, strong domain randomization, MLLM integration | Excellent for generating diverse synthetic data, sim-to-real transfer strategies |
| General Simulation (e.g., Bullet/Isaac Gym) | Synthetic | Diverse household/industrial objects (user-defined) | User-defined | User-defined | Flexible environment for custom scenarios, physics engine integration | Foundation for custom synthetic data generation and testing, rapid prototyping |

**4. Leveraging Your Dataset for Advanced Research**

The planned data collection, featuring a top-down PhotoNeo RGB-D camera observing up to five known items from distinct categories in cluttered bins, offers unique opportunities for advancing robotic perception through "cross-bin inference" and strategic synthetic data augmentation.

**4.1. Cross-Bin Inference Opportunities**

The concept of "cross-bin inference"—utilizing multiple partial views of the same known item observed across different bins to improve learning or recognition—is a unique and powerful extension of multi-view learning. Unlike traditional Multi-View Stereo (MVS) methods, which typically assume dense, overlapping views for 3D reconstruction 30, this scenario involves disparate, partial views of the same known object observed at different times and in different spatial configurations. This shifts the challenge from precise geometric registration of simultaneous dense views to robust feature matching and probabilistic fusion across sparse, potentially heavily occluded observations.

**4.1.1. Multi-View Learning from Disparate Observations**

Each bin presents a partial observation of the items within it. When the same item appears in multiple bins throughout the day, these seemingly independent observations can be aggregated to form a more complete understanding of the object's 3D properties. The principle of aggregating information from multiple viewpoints, as seen in MVS, remains highly relevant, but the application here is distinct. It is not about reconstructing a single scene from multiple camera angles at one moment, but rather about building a richer, more robust model of *each known object* by accumulating fragmented visual evidence over time and across different instances of bins. This could involve techniques that infer "physible geometry" from observed motion and interactions, even under heavy occlusion, to supplement visible geometry.31 Similarly, inferring superquadric representations from locally visible geometry without requiring multi-camera views or prior object knowledge, as demonstrated by RGBSQGrasp, is highly applicable for leveraging partial views.3

**4.1.2. Implicit Shape Completion and Pose Refinement**

The partial observations from a fixed top-down camera in one bin can be complemented by partial observations of the same known item in other bins. This can be conceptualized as solving a "system of constraints," where each partial view provides additional information to refine a complete 3D model or 6D pose estimate for that specific item. For instance, if a bottle is occluded from the top-down view in one bin, a different orientation of the same bottle in another bin might reveal the occluded features. Instead of attempting explicit mesh reconstruction, an approach could explore implicit neural representations, such as NeRF-like models 17, for each known item category. These models could be trained to learn a canonical 3D shape from diverse partial views of the same item collected across different bins, effectively leveraging the cross-bin inference as a form of multi-view learning over time and space. This approach sidesteps the difficulty of explicit 3D reconstruction from heavily occluded, single-view data, offering a more robust path to 3D understanding for grasp prediction in challenging bin-picking scenarios.

**4.1.3. Leveraging Semantic Priors for Robust Matching**

Given that the items are "known" and belong to "distinct object categories," semantic priors can play a significant role in cross-bin inference. Vision-Language Models (VLMs), as seen in ObjectVLA 10, can establish robust links between visual features and semantic categories. This mechanism can be adapted to reinforce the recognition of known objects from partial views, even for texture-less items, by associating fragmented visual data with their categorical identity. This semantic grounding can significantly improve the accuracy of matching partial views to the correct known item in the catalogue, especially when visual information is limited.

**4.2. Synthetic Data Augmentation Strategies**

The ability to generate synthetic data or augmented views is a powerful asset for supplementing the real-world dataset, particularly in addressing the challenges of occlusion and clutter.

**4.2.1. Generating Augmented Views for Occlusion Handling**

Synthetic data generation can be strategically focused on augmenting the specific types of challenges encountered in the real warehouse, particularly severe occlusion and partial views. This involves rendering known items with varying degrees of occlusion, different lighting conditions, and diverse backgrounds, specifically targeting the limitations observed in the real-world data. Frameworks like RoboTwin 2.0 provide a robust blueprint for generating effective synthetic data through structured domain randomization, including variations in clutter, lighting, and background.23 This allows for the creation of synthetic scenarios that systematically vary occlusion patterns and object orientations within the bin, effectively filling the gaps in real-world observations and enabling models to "imagine" or infer occluded parts.

**4.2.2. Sim-to-Real Transfer Techniques for Data Augmentation**

To ensure that synthetically generated data is photorealistic and effectively bridges the reality gap, advanced sim-to-real techniques are indispensable. Randomized-to-Canonical Adaptation Networks (RCANs) can translate randomized simulated images into non-randomized canonical versions, which then allows real-world images to be transformed into simulated equivalents for consistent training.27 Similarly, Generative Adversarial Networks (GANs), such as CycleGAN, can improve the photorealism of synthetic images, making them more closely resemble real-world captures.28 Employing these methods is crucial for ensuring that models trained on synthetic data perform robustly on the real-world data collected from the production warehouse.

**5. Novel and Under-Explored MSc Thesis Directions**

The following thesis directions are tailored to a 6-month MSc scope, leveraging the unique context of the planned data collection (PhotoNeo RGB-D, top-down, known items, cross-bin inference, synthetic data capability) and addressing identified research gaps.

**5.1. Thesis Direction 1: Probabilistic 3D Shape Reconstruction and Pose Estimation via Cross-Bin Multi-View Fusion**

**Key Research Question:** How can partial, occluded RGB-D observations of known items across multiple bins be probabilistically fused to reconstruct accurate 3D shapes and estimate 6D poses, particularly for challenging industrial objects, within a fixed top-down camera setup?

**Problem Context:** The fixed top-down camera limits the completeness of single-view observations, especially with partial occlusion and overlapping items. While the items are known, industrial objects (e.g., texture-less, metallic, symmetrical) pose significant challenges for 3D understanding. The cross-bin inference capability provides multiple, albeit disparate, partial views of the same object over time.

**Proposed Methodology:**

* **Data Preprocessing:** Convert the PhotoNeo RGB-D data into dense point clouds, building on the observation that RGB-Points input offers superior precision for grasp region prediction.2 Develop or fine-tune a robust instance segmentation model to segment individual item instances, even when partially visible, potentially using synthetic data for augmentation.
* **Implicit 3D Representation Learning:** Instead of aiming for explicit mesh reconstruction, explore implicit neural representations, such as NeRF-like models 17, for each

*known item category*. These models would be trained to learn a canonical 3D shape from diverse *partial views* of the same item collected across different bins. This approach frames "cross-bin inference" as a form of multi-view learning distributed over time and space, allowing the model to learn a more complete object understanding from fragmented observations.

* **Probabilistic Fusion & Pose Estimation:** Develop a probabilistic framework, such as a Bayesian network or a graph neural network, that integrates features from multiple partial views of the same item (from different bins). This framework would iteratively refine its 3D shape and 6D pose estimate, treating each partial observation as a constraint in a larger system, aligning with the user's analogy.
* **Synthetic Data Integration:** Generate synthetic data using domain randomization techniques (e.g., following RoboTwin 2.0 principles 23) with controlled occlusion patterns and varying object orientations. This synthetic data would specifically augment the real-world dataset for training the implicit 3D representation and the probabilistic fusion model. Employ sim-to-real techniques like RCANs 27 or GANs 28 to ensure the synthetic data effectively bridges the reality gap.

**Expected Outcomes:** This research is expected to yield improved 3D reconstruction accuracy and 6D pose estimation for highly occluded industrial items compared to methods relying solely on single-view observations. The outcome would be a robust framework capable of leveraging sparse, disparate multi-view data collected over time.

**Scope/Feasibility (6 months):** The project can be scoped by focusing on a subset of the known items (e.g., 2-3 challenging categories) to demonstrate proof-of-concept. The primary task would be to develop and evaluate the implicit 3D representation and the probabilistic fusion module. Benchmarking against a strong single-view baseline would be essential. The existing PhotoNeo data collection and synthetic data generation capabilities are key enablers for this direction.

**5.2. Thesis Direction 2: Few-Shot Grasp Success Prediction for Occluded Industrial Items via Object-Centric Features**

**Key Research Question:** How can a robotic system learn to predict optimal grasp poses and their success probabilities for partially occluded, known industrial items in cluttered bins with only a few real-world demonstrations, leveraging object-centric features from top-down RGB-D input?

**Problem Context:** Accurate estimation of grasp probability is crucial for efficient bin-picking, but real-world grasp trials are costly and time-consuming. Industrial items present unique challenges for grasping due to their texture-less, symmetrical, and potentially reflective properties, often found in dense clutter.19

**Proposed Methodology:**

* **Object-Centric Feature Extraction:** Implement a perception pipeline that processes PhotoNeo RGB-D data, ideally converting it to RGB-Points format for enhanced precision.2 The pipeline would extract robust object-centric features, such as Contact-centric Grasp Representations (CGRs) from methods like AnyDexGrasp.13 This approach focuses on local geometry relevant for grasping, rather than requiring full object pose estimation. Integrate Object Depth Mapping 5 for improved handling of occlusion by simplifying the object's spatial representation.
* **Few-Shot Grasp Prediction Model:** Adapt a few-shot learning approach, such as ControlVLA's object-centric adaptation strategy 7, to predict grasp quality or success probability. This would involve pre-training a general policy on a large synthetic dataset (e.g., using RoboTwin 2.0 23 with extensive domain randomization to cover various clutter and lighting conditions).
* **Targeted Fine-Tuning with Real Data:** Fine-tune the pre-trained model using a very limited number of real-world grasp attempts from the collected dataset. The focus would be on demonstrating how minimal real-world interaction can significantly boost performance on the specific known industrial items.
* **Evaluation:** Measure the grasp success rate on novel configurations of known items in cluttered bins. Quantify the data efficiency by determining how few real-world demonstrations are needed to achieve a target success rate.

**Expected Outcomes:** The research is anticipated to produce a highly data-efficient grasp prediction model that performs robustly on occluded industrial items. It would demonstrate the significant advantages of combining extensive pre-training on synthetic data with minimal, targeted real-world fine-tuning.

**Scope/Feasibility (6 months):** This project can be scoped by focusing on a specific gripper type (e.g., suction or parallel-jaw, relevant to the warehouse context) and selecting a challenging subset of the known items. The primary task is to develop and evaluate the few-shot adaptation strategy for grasp success prediction, assuming that object segmentation or detection is handled by a separate module.

**5.3. Thesis Direction 3: Learning Robust Semantic-Geometric Features for Industrial Object Recognition Under Extreme Occlusion**

**Key Research Question:** How can a perception system learn robust, discriminative semantic-geometric features from top-down RGB-D data to accurately recognize and localize known, texture-less, and highly occluded industrial items in dense clutter, leveraging cross-bin partial views?

**Problem Context:** The "unsolved reality gap" for industrial objects, characterized by texture-less, metallic, and symmetrical properties, combined with dense clutter and severe occlusion 19, presents a formidable challenge for object recognition. The user's scenario involves known items from distinct categories, and cross-bin inference offers multiple partial views of the same object.

**Proposed Methodology:**

* **Multi-Modal Feature Fusion Network:** Design a deep learning architecture that effectively fuses RGB visual cues with 3D geometric information (point clouds derived from the PhotoNeo camera). Explore advanced fusion strategies beyond simple concatenation, potentially incorporating attention mechanisms 34 to dynamically weigh the importance of different features (color, texture, depth, shape) based on context.
* **Semantic-Geometric Feature Learning:** Train the network to learn features that are highly discriminative for each known object category, even from partial views. This could involve integrating semantic priors from Vision-Language Models (VLMs), as demonstrated by ObjectVLA 10, which can effectively link visual features to object names or categories. This is particularly valuable for texture-less items where traditional visual features are scarce.
* **Cross-Bin Consistency Loss:** Introduce a novel loss function that explicitly encourages consistency in the learned features or object identities for the same item observed across different bins. This loss would leverage the "cross-bin inference" concept to reinforce robust feature learning from fragmented and partial views, effectively treating multiple partial observations as a collective source of information for a single object's identity.
* **Synthetic Data for Feature Robustness:** Systematically generate synthetic data with varying occlusion levels, lighting conditions, and clutter densities (following principles from RoboTwin 2.0 23). This synthetic data would be crucial for training the network to learn features that are invariant to these real-world variations and robust to extreme occlusion.

**Expected Outcomes:** This research aims to achieve significantly improved object recognition and localization accuracy for industrial items in cluttered, occluded bins. The outcome would be a novel feature learning approach that effectively combines semantic and geometric information from sparse, partial observations, thereby bridging the current performance gap in industrial settings.

**Scope/Feasibility (6 months):** The project can be scoped to focus primarily on the object recognition and localization components, without necessarily integrating the full pick-and-place pipeline. Performance would be evaluated using standard metrics such as mean Average Precision (mAP) for detection and 6D pose error. A representative subset of known items should be selected to demonstrate the efficacy of the proposed feature learning approach.

**Table 4: Proposed MSc Thesis Directions with Key Research Questions, Methodologies, and Outcomes**

| Thesis Title | Key Research Question | Proposed Methodology | Relevance to User's Context | Expected Outcomes | Scope/Feasibility (6 months) |
| --- | --- | --- | --- | --- | --- |
| Probabilistic 3D Shape Reconstruction and Pose Estimation via Cross-Bin Multi-View Fusion | How can partial, occluded RGB-D observations of known items across multiple bins be probabilistically fused to reconstruct accurate 3D shapes and estimate 6D poses, particularly for challenging industrial objects, within a fixed top-down camera setup? | Convert RGB-D to point clouds; learn implicit 3D representations (NeRF-like) for known items from diverse partial views across bins; develop probabilistic fusion (Bayesian/GNN) for iterative refinement of 3D shape/6D pose; integrate synthetic data with domain randomization and sim-to-real techniques. | Directly leverages PhotoNeo RGB-D, top-down view, known items, and cross-bin inference for challenging industrial objects. | Improved 3D reconstruction accuracy and 6D pose estimation for highly occluded industrial items; robust framework for sparse, disparate multi-view data. | Focus on 2-3 challenging item categories; develop implicit 3D representation and fusion module; benchmark against single-view baseline. |
| Few-Shot Grasp Success Prediction for Occluded Industrial Items via Object-Centric Features | How can a robotic system learn to predict optimal grasp poses and their success probabilities for partially occluded, known industrial items in cluttered bins with only a few real-world demonstrations, leveraging object-centric features from top-down RGB-D input? | Extract object-centric features (e.g., CGRs from RGB-Points); adapt few-shot learning (e.g., ControlVLA) for grasp quality prediction; pre-train on large synthetic dataset (RoboTwin 2.0) with domain randomization; fine-tune with minimal real-world grasp attempts. | Addresses grasp probability for known industrial items; uses PhotoNeo RGB-D; designed for limited real-world demonstrations; directly applicable to warehouse pick-and-place. | Highly data-efficient grasp prediction model robust to occluded industrial items; demonstrates power of pre-training + minimal real-world fine-tuning. | Focus on a specific gripper type; select a challenging subset of known items; develop/evaluate few-shot adaptation for grasp success prediction. |
| Learning Robust Semantic-Geometric Features for Industrial Object Recognition Under Extreme Occlusion | How can a perception system learn robust, discriminative semantic-geometric features from top-down RGB-D data to accurately recognize and localize known, texture-less, and highly occluded industrial items in dense clutter, leveraging cross-bin partial views? | Design multi-modal feature fusion network (RGB + point clouds, with attention); train for discriminative semantic-geometric features from partial views; incorporate semantic priors (VLMs); introduce cross-bin consistency loss; use synthetic data with varied occlusion/clutter for robustness. | Directly tackles the "unsolved reality gap" for industrial objects; leverages PhotoNeo RGB-D and cross-bin inference for known, distinct categories. | Significantly improved object recognition and localization accuracy for industrial items in cluttered, occluded bins; novel feature learning combining semantic and geometric information from sparse observations. | Focus on object recognition/localization; evaluate using standard metrics (mAP, 6D pose error); select representative subset of known items. |

**6. Conclusion and Outlook**

**6.1. Synthesizing Key Findings and Contributions**

The journey towards fully autonomous robotic bin-picking in warehouse logistics is complex, yet critical for meeting the escalating demands of modern supply chains. This report has synthesized the current state of robotic perception, highlighting the indispensable role of RGB-D cameras for acquiring both visual and geometric information in cluttered environments. It has been established that while top-down views offer practical advantages, they necessitate highly robust perception algorithms capable of inferring complete object information from partial and occluded observations. The superiority of RGB-Points as an input modality for grasp prediction and the effectiveness of object-centric representations in handling occlusion have been underscored.

Furthermore, the analysis emphasized the growing importance of data-efficient learning paradigms, such as few-shot and one-shot methods, which allow robots to generalize from minimal demonstrations. Frameworks like ControlVLA and ObjectVLA demonstrate how pre-trained Vision-Language Models and object-centric adaptation can significantly reduce the need for extensive real-world data, enabling rapid deployment and generalization to novel objects. Similarly, two-stage learning approaches, exemplified by AnyDexGrasp, achieve human-level learning efficiency by decoupling hand-agnostic representation learning from hand-dependent grasp decision-making.

A significant finding from the review of existing datasets, particularly XYZ-IBD, is the pronounced and currently unresolved "reality gap" for industrial bin-picking. The unique challenges posed by texture-less, metallic, and symmetrical objects in dense clutter highlight that traditional academic benchmarks often do not reflect authentic industrial conditions. This underscores the strategic importance of collecting real-world data, as planned, to specifically address this gap. This real-world data, when combined with advanced synthetic data generation techniques (e.g., RoboTwin 2.0) and sophisticated sim-to-real transfer methods (e.g., RCANs, GANs), forms a powerful hybrid data strategy that maximizes data diversity while minimizing annotation burden.

The proposed MSc thesis directions directly leverage these insights, offering focused research avenues that can significantly advance the state-of-the-art. By exploring probabilistic 3D shape reconstruction via cross-bin multi-view fusion, few-shot grasp success prediction with object-centric features, and robust semantic-geometric feature learning under extreme occlusion, these projects aim to develop perception systems that are more accurate, data-efficient, and resilient to the complexities of real-world industrial environments.

**6.2. Future Directions and Broader Impact**

The research outlined herein represents a significant step towards developing more adaptable and efficient robotic systems for industrial logistics. Beyond the immediate scope of a 6-month MSc thesis, several promising future directions emerge. Further work could explore the integration of active perception strategies, such as limited robot movement or bin tilting, to acquire more informative views, even within a largely fixed top-down setup. Investigating the handling of deformable objects, which are increasingly common in e-commerce, presents another complex challenge that current rigid-object-focused methods do not fully address. The incorporation of haptic feedback during grasping could provide crucial information for refining grasp stability and success probability, especially for items with ambiguous visual properties.

The broader impact of this research extends far beyond individual warehouses. By developing robust and data-efficient bin-picking solutions, this work contributes to enhancing supply chain resilience, reducing operational costs, and improving worker safety by automating hazardous or monotonous tasks. The advancements in few-shot learning, cross-bin inference, and robust feature extraction for challenging industrial materials will contribute to the development of more general-purpose robotic manipulation capabilities, paving the way for autonomous systems that can adapt quickly to new products and dynamic environments. Ultimately, this research fosters the creation of intelligent robotic agents that can seamlessly integrate into complex industrial ecosystems, driving the next generation of automation and machine support.

**References**

* 5

<https://arxiv.org/html/2504.20969v2>

* 2

<https://www.researchgate.net/publication/332463067_Combining_RGB_and_Points_to_Predict_Grasping_Region_for_Robotic_Bin-Picking>

* 7

<https://arxiv.org/html/2506.16211v1>

* 4

<https://www.researchgate.net/publication/380945031_Robot_Grasp_in_Cluttered_Scene_Using_a_Multi-Stage_Deep_Learning_Model>

* 18

<https://arxiv.org/abs/2210.12126>

* 23

<https://arxiv.org/html/2506.18088v1>

* 19

<https://arxiv.org/html/2506.00599v1>

* 10

<https://arxiv.org/html/2502.19250v1>

* 30

<https://openaccess.thecvf.com/content/CVPR2023/papers/Zhang_GeoMVSNet_Learning_Multi-View_Stereo_With_Geometry_Perception_CVPR_2023_paper.pdf>

* 27

<https://www.researchgate.net/publication/329771897_Sim-to-Real_via_Sim-to-Sim_Data-efficient_Robotic_Grasping_via_Randomized-to-Canonical_Adaptation_Networks>

* 28

<https://www.edi.lv/wp-content/uploads/2022/12/IMOCO_paper_for_EFTA2022_workshop.pdf>

* 6

<https://www.edi.lv/wp-content/uploads/2022/12/IMOCO_paper_for_EFTA2022_workshop.pdf>

* 5

<https://arxiv.org/html/2504.20969v2>

* 9

<https://www.researchgate.net/publication/392918257_ControlVLA_Few-shot_Object-centric_Adaptation_for_Pre-trained_Vision-Language-Action_Models>

* 8

<https://arxiv.org/abs/2506.16211>

* 34

<https://www.researchgate.net/publication/393573331_Improving_3D_Object_Detection_in_Neural_Radiance_Fields_With_Channel_Attention>

* 35

<https://research.nvidia.com/labs/lpr/publication/blukis2023oneshot/>

* 23

<https://arxiv.org/html/2506.18088v1>

* 24

<https://arxiv.org/abs/2506.18088>

* 20

<https://arxiv.org/pdf/2506.00599>?

* 21

<https://arxiv.org/abs/2506.00599>

* 10

<https://arxiv.org/html/2502.19250v1>

* 11

<https://www.researchgate.net/publication/389392635_ObjectVLA_End-to-End_Open-World_Object_Manipulation_Without_Demonstration>

* 29

<https://ar5iv.labs.arxiv.org/html/1812.07252>

* 27

<https://www.researchgate.net/publication/329771897_Sim-to-Real_via_Sim-to-Sim_Data-efficient_Robotic_Grasping_via_Randomized-to-Canonical_Adaptation_Networks>

* 9

<https://www.researchgate.net/publication/392918257_ControlVLA_Few-shot_Object-centric_Adaptation_for_Pre-trained_Vision-Language-Action_Models>

* 7

<https://arxiv.org/html/2506.16211v1>

* 34

<https://www.researchgate.net/publication/393573331_Improving_3D_Object_Detection_in_Neural_Radiance_Fields_With_Channel_Attention>

* 35

<https://research.nvidia.com/labs/lpr/publication/blukis2023oneshot/>

* 25

<https://www.researchgate.net/publication/392941279_RoboTwin_20_A_Scalable_Data_Generator_and_Benchmark_with_Strong_Domain_Randomization_for_Robust_Bimanual_Robotic_Manipulation>

* 23

<https://arxiv.org/html/2506.18088v1>

* 20

<https://arxiv.org/pdf/2506.00599>?

* 22

<https://www.researchgate.net/publication/311609088_Recovering_6D_Object_Pose_and_Predicting_Next-Best-View_in_the_Crowd>

* 10

<https://arxiv.org/html/2502.19250v1>

* 11

<https://www.researchgate.net/publication/389392635_ObjectVLA_End-to-End_Open-World_Object_Manipulation_Without_Demonstration>

* 13

<https://arxiv.org/html/2502.16420v1>

* 9

<https://www.researchgate.net/publication/392918257_ControlVLA_Few-shot_Object-centric_Adaptation_for_Pre-trained_Vision-Language-Action_Models>

* 8

<https://arxiv.org/abs/2506.16211>

* 35

<https://research.nvidia.com/labs/lpr/publication/blukis2023oneshot/>

* 32

<https://proceedings.neurips.cc/paper_files/paper/2023/file/525d24400247f884c3419b0b7b1c4829-Paper-Conference.pdf>

* 23

<https://arxiv.org/html/2506.18088v1>

* 24

<https://arxiv.org/abs/2506.18088>

* 20

<https://arxiv.org/pdf/2506.00599>?

* 21

<https://arxiv.org/abs/2506.00599>

* 10

<https://arxiv.org/html/2502.19250v1>

* 11

<https://www.researchgate.net/publication/389392635_ObjectVLA_End-to-End_Open-World_Object_Manipulation_Without_Demonstration>

* 14

<https://www.researchgate.net/publication/389314706_AnyDexGrasp_General_Dexterous_Grasping_for_Different_Hands_with_Human-level_Learning_Efficiency>

* 13

<https://arxiv.org/html/2502.16420v1>

* 9

<https://www.researchgate.net/publication/392918257_ControlVLA_Few-shot_Object-centric_Adaptation_for_Pre-trained_Vision-Language-Action_Models>

* 8

<https://arxiv.org/abs/2506.16211>

* 36

<https://arxiv.org/pdf/2502.13335>?

* 34

<https://www.researchgate.net/publication/393573331_Improving_3D_Object_Detection_in_Neural_Radiance_Fields_With_Channel_Attention>

* 23

<https://arxiv.org/html/2506.18088v1>

* 24

<https://arxiv.org/abs/2506.18088>

* 10

<https://arxiv.org/html/2502.19250v1>

* 11

<https://www.researchgate.net/publication/389392635_ObjectVLA_End-to-End_Open-World_Object_Manipulation_Without_Demonstration>

* 14

<https://www.researchgate.net/publication/389314706_AnyDexGrasp_General_Dexterous_Grasping_for_Different_Hands_with_Human-level_Learning_Efficiency>

* 13

<https://arxiv.org/html/2502.16420v1>

* 9

<https://www.researchgate.net/publication/392918257_ControlVLA_Few-shot_Object-centric_Adaptation_for_Pre-trained_Vision-Language-Action_Models>

* 8

<https://arxiv.org/abs/2506.16211>

* 37

<https://arxiv.org/pdf/2503.21732>

* 34

<https://www.researchgate.net/publication/393573331_Improving_3D_Object_Detection_in_Neural_Radiance_Fields_With_Channel_Attention>

* 23

<https://arxiv.org/html/2506.18088v1>

* 24

<https://arxiv.org/abs/2506.18088>

* 10

<https://arxiv.org/html/2502.19250v1>

* 11

<https://www.researchgate.net/publication/389392635_ObjectVLA_End-to-End_Open-World_Object_Manipulation_Without_Demonstration>

* 14

<https://www.researchgate.net/publication/389314706_AnyDexGrasp_General_Dexterous_Grasping_for_Different_Hands_with_Human-level_Learning_Efficiency>

* 13

<https://arxiv.org/html/2502.16420v1>

* 9

<https://www.researchgate.net/publication/392918257_ControlVLA_Few-shot_Object-centric_Adaptation_for_Pre-trained_Vision-Language_Action_Models>

* 7

<https://arxiv.org/html/2506.16211v1>

* 17

<https://dtransposed.github.io/blog/2022/08/06/NeRF/>

* 33

[https://openreview.net/forum?id=AGG1zlrrMw¬eId=AGG1zlrrMw](https://openreview.net/forum?id=AGG1zlrrMw&noteId=AGG1zlrrMw)

* 26

<https://www.themoonlight.io/en/review/robotwin-20-a-scalable-data-generator-and-benchmark-with-strong-domain-randomization-for-robust-bimanual-robotic-manipulation>

* 23

<https://arxiv.org/html/2506.18088v1>

* 10

<https://arxiv.org/html/2502.19250v1>

* 11

<https://www.researchgate.net/publication/389392635_ObjectVLA_End-to-End_Open-World_Object_Manipulation_Without_Demonstration>

* 15

<https://arxiv.org/abs/2502.16420>

* 16

<https://graspnet.net/anydexgrasp/assets/files/AnyDexGrasp.pdf>

* 9

<https://www.researchgate.net/publication/392918257_ControlVLA_Few-shot_Object-centric_Adaptation_for_Pre-trained_Vision-Language_Action_Models>

* 7

<https://arxiv.org/html/2506.16211v1>

* 17

<https://dtransposed.github.io/blog/2022/08/06/NeRF/>

* 33

[https://openreview.net/forum?id=AGG1zlrrMw¬eId=AGG1zlrrMw](https://openreview.net/forum?id=AGG1zlrrMw&noteId=AGG1zlrrMw)

* 26

<https://www.themoonlight.io/en/review/robotwin-20-a-scalable-data-generator-and-benchmark-with-strong-domain-randomization-for-robust-bimanual-robotic-manipulation>

* 23

<https://arxiv.org/html/2506.18088v1>

* 10

<https://arxiv.org/html/2502.19250v1>

* 12

<https://www.themoonlight.io/en/review/objectvla-end-to-end-open-world-object-manipulation-without-demonstration>

* 15

<https://arxiv.org/abs/2502.16420>

* 16

<https://graspnet.net/anydexgrasp/assets/files/AnyDexGrasp.pdf>

* 9

<https://www.researchgate.net/publication/392918257_ControlVLA_Few-shot_Object-centric_Adaptation_for_Pre-trained_Vision-Language_Action_Models>

* 7

<https://arxiv.org/html/2506.16211v1>

* 17

<https://dtransposed.github.io/blog/2022/08/06/NeRF/>

* 33

[https://openreview.net/forum?id=AGG1zlrrMw¬eId=AGG1zlrrMw](https://openreview.net/forum?id=AGG1zlrrMw&noteId=AGG1zlrrMw)

* 26

<https://www.themoonlight.io/en/review/robotwin-20-a-scalable-data-generator-and-benchmark-with-strong_domain_randomization_for_robust_bimanual_robotic_manipulation>

* 23

<https://arxiv.org/html/2506.18088v1>

* 10

<https://arxiv.org/html/2502.19250v1>

* 11

<https://www.researchgate.net/publication/389392635_ObjectVLA_End-to-End_Open-World_Object_Manipulation_Without_Demonstration>

* 15

<https://arxiv.org/abs/2502.16420>

* 16

<https://graspnet.net/anydexgrasp/assets/files/AnyDexGrasp.pdf>

* 31

<https://roboticsconference.org/program/papers/34/>

* 3

<https://arxiv.org/html/2503.02387v1>

* 1

<https://www.researchgate.net/publication/254098851_Fast_object_localization_and_pose_estimation_in_heavy_clutter_for_robotic_bin_picking_The_International_Journal_of_Robotics_Research_318_951-973>

* 31

<https://roboticsconference.org/program/papers/34/>

* 3

<https://arxiv.org/html/2503.02387v1>

* 2

<https://www.researchgate.net/publication/332463067_Combining_RGB_and_Points_to_Predict_Grasping_Region_for_Robotic_Bin-Picking>

* 28

<https://www.edi.lv/wp-content/uploads/2022/12/IMOCO_paper_for_EFTA2022_workshop.pdf>

* 30

<https://openaccess.thecvf.com/content/CVPR2023/papers/Zhang_GeoMVSNet_Learning_Multi-View_Stereo_With_Geometry_Perception_CVPR_2023_paper.pdf>

* 29

<https://ar5iv.labs.arxiv.org/html/1812.07252>

* 29

<https://ar5iv.labs.arxiv.org/html/1812.07252>

* 20

<https://arxiv.org/pdf/2506.00599>?

* 20

<https://arxiv.org/pdf/2506.00599>?

* 15

<https://arxiv.org/abs/2502.16420>

* 16

<https://graspnet.net/anydexgrasp/assets/files/AnyDexGrasp.pdf>