MetaGraspNetV2: All-in-One Dataset Enabling Fast and Reliable Robotic Bin Picking via Object Relationship Reasoning and Dexterous Grasping

Maximilian Gilles, Yuhao Chen, *Member, IEEE*, Emily Zhixuan Zeng, *Member, IEEE*, Yifan Wu, Kai Furmans, *Member, IEEE*, Alexander Wong, and Rania Rayyes

***Abstract*— Grasping unknown objects in unstructured environments is one of the most challenging and demanding tasks for robotic bin picking systems. Developing a holistic approach is crucial to building such dexterous bin picking systems to meet practical requirements on speed, cost and reliability. Proposed datasets so far focus only on challenging sub-problems and are therefore limited in their ability to leverage the complementary relationship between individual tasks. In this paper, we tackle this holistic data challenge and design MetaGraspNetV2, an all-in-one bin picking dataset consisting of (i) a photo-realistic dataset with over 296k images, which has been created through physics-based metaverse synthesis; and (ii) a real-world test dataset with 3.2k images featuring task-specific difficulty levels. Both datasets provide full annotations for amodal panoptic segmentation, object relationship detection, occlusion reasoning, 6-DoF pose estimation, and grasp detection for a parallel-jaw as well as a vacuum gripper. Extensive experiments demonstrate that our dataset outperforms state-of-the-art datasets in object detection, instance segmentation, amodal detection, parallel-jaw grasping, and vacuum grasping. Furthermore, leveraging the potential of our data for building holistic perception systems, we propose a single-shot-multi-pick (SSMP) grasping policy for scene understanding accelerated fast picking in high clutter. SSMP reasons about suitable manipulation orders for blindly picking multiple items given a single image acquisition. Physical robot experiments demonstrate that SSMP effectively speeds up cycle times through reducing image acquisitions by more than 47% while providing better grasp performance compared to state-of-the-art bin picking methods.**

Manuscript received 4 September 2023; accepted 25 October 2023. Date of publication 6 November 2023; date of current version 8 August 2024. This article was recommended for publication by Associate Editor G. Palli and Editor J. Yi upon evaluation of the reviewers' comments. This work was supported in part by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) under Grant 01MJ21007B and in part by the National Research Council Canada (NRC). The work of Rania Rayyes was supported by the Baden-Wurttemberg Ministry of Science, Research and the Arts Within Innovations Campus Mobilität der Zukunft (ICM). *(Corresponding author: Maximilian Gilles.)*

Maximilian Gilles, Kai Furmans, and Rania Rayyes are with the Institute of Material Handling and Logistics, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany (e-mail: maximilian.gilles@kit.edu).

Yuhao Chen, Emily Zhixuan Zeng, Yifan Wu, and Alexander Wong are with the Image Processing Laboratory, University of Waterloo (UW), Waterloo, ON N2L 3G1, Canada.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TASE.2023.3328964.

Digital Object Identifier 10.1109/TASE.2023.3328964

***Note to Practitioners*—In robotic bin picking, most proposed methods and datasets focus on solving only one aspect of the grasping task, such as grasp point detection, object detection, or relationship reasoning. They do not address practical aspects such as the widespread use of vacuum grasp technology or the need for short cycle times. In practice, however, efficient bin picking solutions often rely on multiple task-specific methods. Hence, having one dataset for a large variety of vision-related tasks in robotic picking reduces data redundancy and enables the development of holistic methods. While deep learning has been proven highly effective for bin picking vision systems, it demands large, high-quality training datasets. Collecting such datasets in the real-world, while assuring label quality and consistency, is prohibitively expensive and time-consuming. To overcome these challenges, we set up a photo-realistic metaverse data generation pipeline and create a large-scale synthetic training dataset. Furthermore, we design a comprehensive real-world dataset for testing. Unlike previously proposed datasets, our datasets provide difficulty levels and annotations in simulation and real-world for a comprehensive list of high-level tasks, including amodal object detection, scene layout reasoning, and grasp detection. In real-world applications, cycle time is a critical factor affecting the productivity and profitability of a robotic system. We tackle time-efficiency through scene understanding and demonstrate the capability of our data regarding holistic system development by proposing a single-shot-multi-pick (SSMP) policy. Our SSMP algorithm, trained exclusively on our synthetic data, distinguishes between uncovered and occluded items, and infers specific manipulation orders to perform multiple blind picks in a single shot. Physical robot experiments show that SSMP was able to reduce image acquisitions by more than 47% without compromising grasp performance. This clearly demonstrates that SSMP, together with our dataset, paves the way for application-oriented research in time-critical bin picking.**

***Index Terms*— Bin picking, dataset, object detection, object relationship reasoning, panoptic amodal segmentation, parallel-jaw grasping, pose estimation, vacuum grasping.**

I. Introduction

IN WAREHOUSING and manufacturing, robotic bin picking is an essential task for order-picking or machine feeding processes. Unstructured bin environments with highly occluded and unknown objects pose major challenges for such autonomous systems, increasing investment costs and compromising performance. To overcome these challenges and accelerate the development of reliable and fast universal

bin picking robot systems, three vision-related key challenges regarding robotic bin picking are specified:

A. Robust Object Detection

Finding objects of interest inside a grasp scene can be formulated as an object detection problem. Although this task is not an essential step for many data-driven grasp detection methods [1], [2], [3], work such as [4] and [5] show that an upstream object recognition can significantly improve the grasping performance. In addition, robust object detection is a preliminary step for high-level scene understanding related tasks such as 6-DoF pose estimation, object relationship detection, and manipulation order reasoning [6], [7], [8]. Robotic grasping sensors are typically color or depth cameras. Combining both modalities is common, as color data captures fine object details but struggles with similarly textured items and lighting variations, while depth data excels at capturing scale-free object geometries but is prone to noise and material properties. The successful use of both types of modality is a non-trivial problem and is actively studied in automation [9]. Another challenge is the strong occlusion and stacking of objects inside a bin, which results in shadow areas and reduced information. Besides, thin and non-convex objects often visually break into multiple parts and are likely to be detected as multiple instances [10]. In order to detect unseen objects or deformable objects, the vision systems needs to understand objects at a basic texture and geometry level [5].

We address the challenge of reliable object detection in bin scenes from a data perspective, providing panoptic segmentation labels as well as object detection specific difficulty scores. Together, our dataset enables effective and customized testing across difficulty levels and sensor modalities.

B. Reliable and Dexterous Grasping Strategies

Various different types of grippers exist to handle a variety of objects, differing in shape, weight, or material. Among the most common in industry are vacuum suction cups and parallel-jaw grippers. Vacuum grippers are widely used owing to their ease of operation and grasp point detection, as well as low cost and compact size at the robot's end effector. Therefore, they are well-suited for bin picking tasks and are preferred in real-world applications. However, when dealing with a high-diverse article spectrum, vacuum grippers may not be able to handle all items due to material or geometric constraints. In such cases, swapping the end effector for an additional parallel gripper can be advantageous, since it has lower demands on the object's material and geometric properties [3], [11]. Regardless of the employed gripper technology, grasping occluded items may result in damage, unintentional double picks or failed attempts due to high torques in the gripper-object contact. Finding reliable grasps for a given grasping scene is not limited to reasoning about the gripper-object interaction in isolation. It also requires to understand the underlying relationship between neighboring objects.

In order to assure reliable grasping strategies, we contribute high-quality grasp annotations, 6-DoF pose annotations, and amodal segmentation masks to provide the necessary information and enrich the dataset for skillful and dexterous manipulation.

C. Fast Cycle Times

When installing picking cells, specific time requirements are defined beforehand. The system's cycle time is a critical factor to meet the return of investment or feeding rates of downstream tasks. One prominent approach to speed up the system is to reduce the idle time of the robot. In practice, this task is often fulfilled by a system integrator optimizing the physical layout and adapting the program sequence with use-case specific solutions. An advanced vision system can also play a significant role in reducing cycle times by, for instance, reasoning about sets of objects that can be picked within a single image acquisition. In order to successively pick objects blindly, the picked objects must not change the grasp scene while picking. While commercial systems can filter detected objects by their distance or assume a semi-structured scene, one-shot multi-item picking in clutter has not received much attention in research yet, despite its high economic potential. Especially in cluttered scenes, more high-level information about the object layout is required to reason about suitable blind picking sequences.

Our dataset with its comprehensive label set, including object relationship graphs and occlusion information, enhances the development of such fast picking systems and enables application in unstructured scenes.

Various datasets have been proposed focusing on individual aspects of bin picking related to the aforementioned challenges, e.g. WISDOM [5], [10] for object segmentation in high clutter, SynPick [12] for pose estimation, REGRAD [13] for object relationship reasoning, UOAIS [14] for amodal segmentation, SuctionNet-1Billion [2] for vacuum grasping, and DexNet 4.0 [3] for ambidextrous grasping. However, in contrast to our proposed dataset, these datasets are capable of addressing only a sub-range of the challenges mentioned above. While holistic system design is critical for real-world robotics in order to meet requirements regarding speed and reliability, individual datasets addressing sub-problems only advance the development of such systems to a limited extend. Accordingly, a more general encompassing dataset approach is needed to overcome the complexities and nuances of automation-specific challenges and to create universal bin picking solutions.

With the MetaGraspNetV2 dataset, a novel all-in-one dataset is designed that aims to unify the most critical aspects of robot vision and bin picking (cf. Fig. 1). Collecting such comprehensive label sets from experiments [15], [16] or manually [11] is too expensive and time prohibitive. Motivated by work demonstrating the generalization capabilities towards real-world data [3], [13], [14], our previous work [17] proposed a data creation pipeline based on metaverse synthesis and contributed two datasets: a large-scale synthetic training dataset and smaller, but comprehensive real-world test dataset.

In this paper, we build on our previously proposed synthetic data generating pipeline (cf. Sec. III) and increase the sample

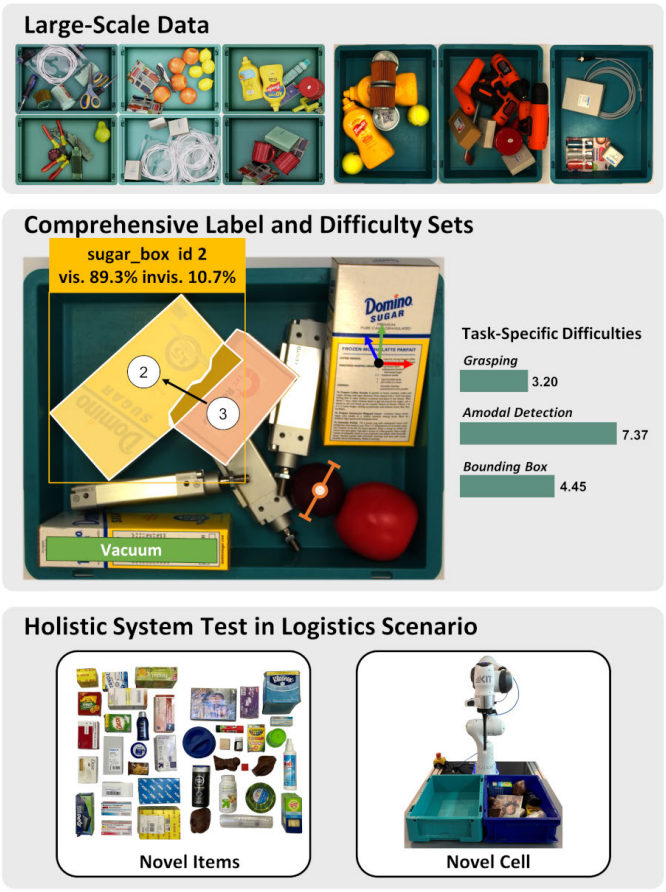


Fig. 1. MetaGraspNetV2 provides a large-scale synthetic and real-world dataset with comprehensive label sets and difficulty levels allowing for holistic system development. Annotations for both datasets include panoptic amodal segmentation masks, object relationship labels, 6-DoF object poses, and ambidextrous grasp labels.

size of of our synthetic dataset (MGN-Sim) and real-world dataset (MGN-Real) by 40%, respectively. To assure performance for real-world problems, a comprehensive and fully-annotated test dataset (MGN-Real) is required. We close the annotation gab between our simulation and real-world data by providing labels for panoptic amodal segmentation and 6-DoF object poses in MGN-Real (cf. Sec. V-A). Task-specific difficulty levels for object detection, amodal detection, and vacuum grasping are designed and allow for efficient performance monitoring (cf. Sec. V-B). Comprehensive label sets enable the development of holistic vision systems. We showcase the capability of our data for time-critical bin picking by developing a novel method for single-shot-multi-pick (SSMP) grasping in Sec. VI. SSMP fuses grasp detection with scene occlusion reasoning, and is able to perform multiple blind picks given a single image acquisition. Physical robot experiments in Section VII-A demonstrate its superiority for vacuum and parallel-jaw grasping in terms of speed and reliability compared to state-of-the-art methods for unseen environments and high clutter. Furthermore, extensive object detection experiments focusing on generalizability show that models trained on our dataset outperform a state-of-the-art dataset for object detection, instance segmentation, and amodal segmentation (cf. Sec. VII-C).

Datasets and code are public available at `https://github.com/maximiliangilles/MetaGraspNet`.

II. Related Work

Datasets for robotic grasping are versatile and differ in many aspects such as scene composition, item diversity, sensor modality, gripper types, and labelled properties. Related work can be categorized based on its sensor modalities in depth-only and photo-realistic RGB-Depth (RGBD) data. While depth-only datasets [1], [3], [25], [26], [27] are sufficient for training grasp detection networks such as [3], [4], [25], [28], and [29], multi-modal RGBD datasets (cf. Table I) are needed for recent sensor fusion approaches [2], [9], [30], [31], [32], or object relationship detection methods [14] combining color and depth information. In the following, we will discuss datasets related to robotic grasping regarding their respective contributions and annotations.

A. Parallel-Jaw Grasping

A top-down grasp with 4-DoF can be represented in image space by an oriented bounding box [18]. Shifting scene and grasp label generation into simulation, [19] can increase the dataset size by a factor up to 50 with regard to dataset size in [18]. As the number of objects and the complexity of the items increases, the more obvious the advantage of 6-DoF grasps over top-down grasps becomes. Annotating grasps in $SE(3)$ with 6-DoF can be tedious. Therefore, often automatic sampling schemes are used, either based on analytical models such as antipodal samplers [13], [21] or simulation environments [3], [26]. For an in-depth overview over sampling methods, we refer to [33]. To provide a broad coverage of grasps even for cluttered scenes, our MetaGraspNetV2 dataset contributes 6-DoF grasp annotations based on antipodal sampling and simulation.

B. Vacuum Grasping

In [11] and [34] vacuum contact points are manually labelled based on human experience. The annotators were experienced users and instructed to annotate pixel regions in bin scenes where a vacuum seal is applicable or not. The manual annotation of vacuum grasps is straightforward and easy to start. However, it is limited in scale and once annotated, adapting an existing dataset to different gripper dimensions is difficult. Besides, human introduced label ambiguities can potentially harm the training process and the performance of the system. Inspired by prior work regarding grasping with two-finger or multi-finger hands, research has been shifted toward focusing on sampling reliable vacuum grasp labels proposing physically-inspired contact force models [1], [2], [25], [35]. Finding seal-tight contact points based on the local suction cup-object interaction results in reliable vacuum grasps and comes with a high potential for energy and cost savings [36], as well as label adaptability [25]. Dex-Net 3.0 [1] models the suction cup as a spring-mass system and predicts well-suited vacuum gripper contact points based on the object's mesh surface. The proposed model is adapted from

TABLE I
OVERVIEW OVER PUBLIC RGBD GRASP DATASETS

Work	Data			Grasp and Scene Labels[a]							
	Item Sets	Sensor	Scene	DoF	PJ[b]	SC[b]	Pose	Seg.	ASeg.	OR	Occl.
[18]	[18]	real	single obj.	4	hand						
[19]	[20]	sim	single obj.	6	sim			✓			
[6]	[6]	real	clutter	4	hand					✓	
[21]	[21], [22]	real	clutter	6	analytic		✓	✓			
[2]	[2], [22]	real	clutter	6		analytic	✓	✓			
[13]	[20]	sim	clutter	6	analytic		✓	✓		✓	
[12]	[23]	sim	clutter	6		heuristic	✓	✓		✓	
[11]	[11]	real	clutter	4	hand	hand					
ours	**ours, [22], [24]**	**sim + real**	**clutter**	**6**	**sim**	**analytic**	✓	✓	✓	✓	✓

[a] DoF: Degree of Freedom Grasp; PJ: Parallel-Jaw Grasp Label Sampling Method; SC: Vacuum Grasp Label Sampling Method; Seg.: Instance Segmentation Mask; ASeg.: Amodal Segmentation Mask; OR: Object Relationship Label; Occl.: Object Occlusion Mask;
[b] hand: manually labelled; sim: physics simulation; analytic: analytical sampling; heuristic: 2D segmentation mask and heuristic

prior work on deformable textile simulation [37] and simplifies the problem by only considering the quasi-static projection of the suction cup onto the object's mesh. Depending on the geometric deformation of the projected springs, a seal is assumed to be applicable or not. SuctionNet-1Billion [2] builds upon [1] simplifying the spring structure and contributes a dataset with real RGBD data and multiple objects per scene. However, building up on prior work [21], their dataset requires manually annotating the objects poses' in the scene which introduces label inaccuracies, limits the items' arrangement, and the overall scene number. Depending on the number of objects and the sample density, simulating the vacuum seal for a single contact point can lead to large computational costs. In [38] and [39], the seal problem is simplified by using geometrically inspired grasp heuristics [40] based on object planarity. By using strong parallelization, we can overcome the computational burden of modelling the vacuum seal with spring-mass models and thus provide realistic and dense vacuum grasp annotations for our synthetic data.

C. Object Detection and Relationship Reasoning

High occlusion and tight packaging present unique challenges to the vision system that are often encountered in automation. With WISDOM [5] a depth-only dataset for instance segmentation of everyday objects in bin scenes is proposed. [10] focuses on highly occluded, thin industrial parts and proposes a method and dataset for robust instance segmentation. Simply inferring grasps without considering the underlying object arrangement can result in unsuccessful grasp attempts or even damaged objects. Recent work addresses this problem by attempting to learn the manipulation order for a picking system [27], [41]. However, currently only a few datasets are available that provide the necessary scene layout information [6], [13] (cf. Table I). VMRD [6] contributes a dataset of over 5k manually annotated scenes with 2D bounding box annotations and relationship labels. The follow-up work, REGRAD [13] uses simulation to increase size and provides 6-DoF parallel-jaw grasps and manipulation order labels. UOAIS [14] reasons about manipulation order predicting visible and invisible masks of items and provides a synthetic dataset UOAIS-SIM for amodal segmentation. To the best of our knowledge, MetaGraspNetV2 is the first dataset providing amodal segmentation masks and object relationship labels in simulation and real-world for highly occluded bin scenes.

D. Pose Estimation and Keypoint Detection

Recent research has increasingly focused on model-free end-to-end grasp detection methods, which do not rely on a 3D model of the object to be grasped. However, for tasks such as packing, machine feeding, or assembling, accurate object pose estimation is still crucial for precise localization and placement of picked items or for downstream manipulation tasks [27], [42], [43]. SynPick [12] contributes a synthetic dataset for 6-DoF pose estimation in dense bin clutter simulating object-gripper interaction. [42] focuses on texture-less, metallic industry objects and sim-to-real transfer. However, both datasets are limited in terms of class diversity and scale or do not provide real-world test data. Regarding pure 6-DoF pose estimation, HOPE [44] and HomebrewDB [45] dataset provide household scenes containing every-day objects arranged in clutter. In MetaGraspNetV2, we provide 6-DoF object pose labels in our simulation and real-world dataset featuring a large-number of different classes with diverse object arrangement and occlusion properties located in an industry typical bin.

III. SYNTHETIC DATA

Our proposed method for synthetic data generation, initially introduced in [17], can be divided into three steps (cf. Fig. 2): A) putting together a diverse object set (cf. Sec. III-A), B) sampling parallel-jaw and vacuum grasp labels for each object individually (cf. Sec. III-B), and C) simulating the physical interaction of objects falling into the grasp scene together with rich annotations for every scene viewpoint (cf. Sec. III-C).

A. Object Dataset

To address challenges posed by unseen objects at test time, we aim for a diverse set of objects, emphasizing versatility in object shape and domain affiliation. The shape of an object

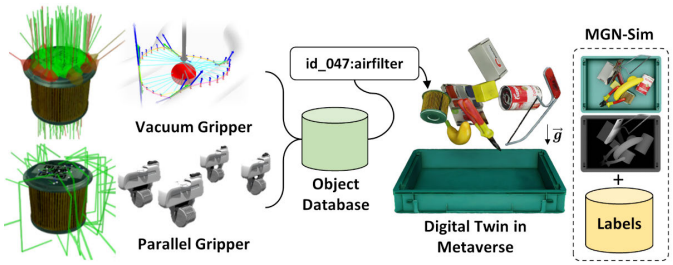


Fig. 2. Synthetic data generation pipeline overview.

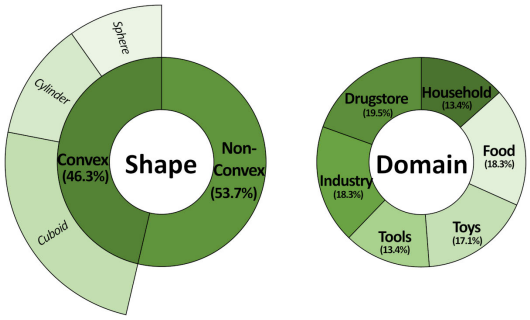


Fig. 3. The object set is evenly balanced between convex (sphere-, cylinder and cuboid-like) and non-convex shaped objects and across domains.

significantly influences its grasping strategy and grasp point availability [46]. Convex shaped objects, such as spheres or cylinders, are easier to grasp with vacuum grippers than with parallel-jaw grippers. While complex, non-convex shaped objects can challenge collision-free grasp detection [47]. In total, our object set consists of 82 high-quality, real-world objects, evenly distributed between convex and non-convex shapes and across domains (cf. Fig. 3). For the composition, we selected 33 meshes from YCB object set [22], 4 from [24], and scanned 36 custom objects using a commercial 3D scanner. For 9 objects scanning was not possible, we remodeled them in CAD software.

B. Grasps Sampling Strategy

In logistics and automation domain, picking robots are usually equipped with vacuum or parallel grippers. In order to deal with a wide range of articles, we provide ambidextrous grasp annotations to train robots with both gripper technologies.

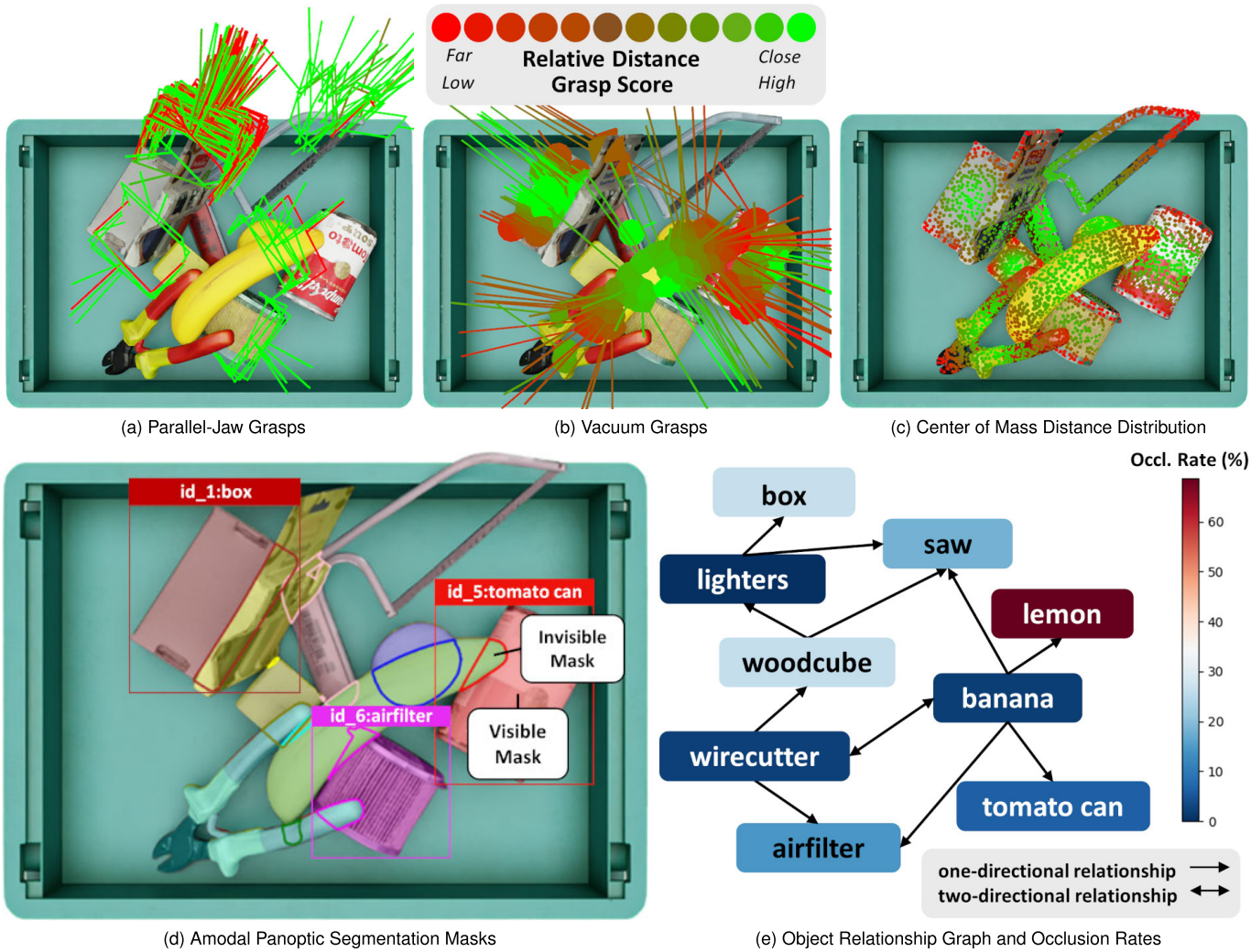
1) Parallel-Jaw Grasps: Parallel-jaw grasp labels G^{pj} are generated combining antipodal sampling strategy and physics simulation [26]. For each object, up to $j = 1 \ldots 5k$ antipodal grasps are generated by sampling finger-object contact points c_i^{pj} evenly distributed over the object's surface. For each contact point c_i^{pj}, $k = 1 \ldots N$, $N = 5$ antipodal [48] contact pairs $\{\tilde{c}_l, \tilde{c}_r\}_{i,k}$ are sampled with deviation in approach direction and translation. The introduced robust antipodal score $s_{\text{antip},i} \in [0, 1]$ for a contact point c_i^{pj} is defined as the number of successful antipodal samples divided by the number of total samples N. To obtain grasp poses in $SE(3)$, for each successful antipodal contact pair, up to $l = 1 \ldots L$, $L = 50$ gripper poses are sampled by rotating it around the fingers' closing direction. A grasp $G_j^{\text{pj}} = G_{i,k,l}^{\text{pj}}$ is considered successful if the gripper model does not collide with the object and it is assigned $s_{\text{anal},j}^{\text{pj}} = s_{\text{antip},i}$. After identifying antipodal contact pairs, each grasp G_j^{pj} is executed multiple times in a physics simulation in IsaacGym [49] using the Franka Hand as gripper reference model. The idea of robustness is employed in simulation by simulating G_j^{pj} with varying mass density factors and friction coefficients. Similar to ACRONYM [26], an upward and shaking gripper movement is performed and a grasp is considered to be successful if the object remains in contact after execution. The robust simulation score $s_{\text{sim},j}^{\text{pj}}$ is defined as the fraction of successful simulated grasps divided by the number of attempts.

2) Vacuum Grasps: An airtight-seal between vacuum cup and object is a necessary condition for vacuum grippers. Though it can be weakened using higher air flow rates, a good seal is important for energy efficiency and grasp stability [36]. To improve the quality of vacuum grasp annotations G^{sc} in simulation, we align with prior work [1], [2], [25] and propose a model able to predict the vacuum sealability of a contact point given the object's 3D mesh. Up to 5k vacuum gripper contact points c_i^{sc} are sampled on the object's surface and checked for seal. For each contact point c_i^{sc}, $k = 1 \ldots N$, $N = 10$ vacuum grasp attempts $c_{i,k}^{\text{sc}}$ are considered with varying approach direction and translation. Similar to the parallel-jaw sampling strategy described above, the robust vacuum grasp sealability score $s_i^{\text{sc}} \in [0, 1]$ is defined as the quotient of successful seals over the number of total samples N. If $s_i^{\text{sc}} > 0$ and the vacuum gripper does not collide with the object in simulation, $G_j^{\text{sc}} = G_i^{\text{sc}}$ is added to the set of successful grasps. Within our vacuum contact model, the projection idea in [1] is adapted due to its universality and efficiency. The suction cup is abstracted as a spring-mass system and its mass points are projected onto the object's surface along the gripper's approaching direction. Based on the projection and the displacement of the springs, forces within the spring system are computed. It is concluded that there is a leak between object and vacuum cup if one of the mass points is not in contact with the object's surface. A more detailed derivation of our model and experimental evaluation can be found in our previous work [17].

C. Scene Generation and Annotations

Instead of manual or semi-automatic methods for generating and labeling grasp scenes [2], [6], [11], [21], we have drawn inspiration from the metaverse trend and create data in NVIDIA Isaac Sim [50]. In a simulated bin picking scenario, objects are randomly sampled and dropped into the bin. The realistic physics-based interaction between the objects k, $k = 1 \ldots N$ and the bin assures that scene layouts are diverse and close to reality. Each scene is captured with alternating lightning conditions from 37 different camera viewpoints arranged in a hemisphere and facing the center of the bin. Path-tracing is used as a rendering setting to capture realistic light and shadow configurations, as well as realistic rendering of materials such as glass, plastic, or metal.

For each viewpoint, all individual objects' parallel $G_{k,j}^{\text{pj}}$ and vacuum suction grasps $G_{k,j}^{\text{sc}}$ (cf. Sec. III-B) are projected into the bin scene and checked for visibility and collision with other



(a) Parallel-Jaw Grasps (b) Vacuum Grasps (c) Center of Mass Distance Distribution

(d) Amodal Panoptic Segmentation Masks (e) Object Relationship Graph and Occlusion Rates

Fig. 4. Our synthetic dataset provides a comprehensive list of ambidextrous grasping (a)-(b) and scene annotations (d)-(e).

objects or with the tote when approaching and performing the grasp (cf. Fig. 4(a)-(b)). A good grasp not only depends on the local object's properties, but is also highly affected by the wrenches applied to the gripper contact. Considering the object's pose and center of mass, the wrench for each vacuum grasp is computed around all three contact axes and scored similar to [2] $s^{\text{sc}}_{\text{sim},j} \in [0, 1]$. Though being implicitly considered in $s^{\text{pj}}_{\text{sim}}$, an explicit soft-finger wrench score [48] $s^{\text{pj}}_{\text{soft},j} \in [0, 1]$ around the gripper's closing direction is also specified for each parallel jaw grasp G^{pj}_j.

Amodal panoptic segmentation masks, instance occlusion rates, and center of mass distribution heat maps are provided for each viewpoint (cf. Fig. 4(c)-(e)). The amodal segmentation mask is defined as a tuple of visible mask and occluded mask for each object instance k assigned to a semantic class in the scene $\{M_{\text{vis}}, M_{\text{occl}}\}_k$. The occlusion score $s_{\text{occl},k} \in [0, 1)$ is defined as the quotient of occluded $M_{\text{occl},k}$ and total object surface area $M_{\text{total},k} = M_{\text{occl},k} \cup M_{\text{vis},k}$. The computation is based on visible object masks by hiding and showing objects in the scene. For the center of mass distribution heatmap, l=1...1000 surface points c_l are sampled for each object and checked for visibility. Mass density is assumed to be the same for all objects. The distance score is defined as $s^{\text{d}}_{l,k} = \{(d_{\max} - d_l)/(d_{\max} - d_{\min})\}_k$, where $d_l = \|c_l - c^{\text{com}}_k\|_2$.

In addition to amodal segmentation masks and occlusion rates, an object relationship graph is proposed to characterize the layout of objects in the scene. For an object pair $(O_k, O_j), k \neq j$ three types of relationships are defined: If O_k is occluding O_j, the relationship (O_k, O_j) is defined as positive and both edges are connected with a directed vertex (k, j), e.g. $(banana, lemon)$ in Fig. 4(e). If O_k is occluding O_j and O_j is occluding O_k, the connection between both objects is defined as interlocked and both edges are connected with a bi-directional vertex $((k, j), (j, k))$ (cf. object pair $(banana, wirecutter)$ in Fig. 4(e)). If O_j and O_j have no direct relationship, the relationship is defined as neutral and no connection between both vertices is added. Starting from the presented relationship graph in Fig. 4(e), one can make the following statements regarding manipulation order: First, assuming that fully visible items should be picked first, we can find unoccluded objects of interest by scanning the graph for object vertices without parents nodes. Secondly, given an object of interest, it is possible to reason about the necessary

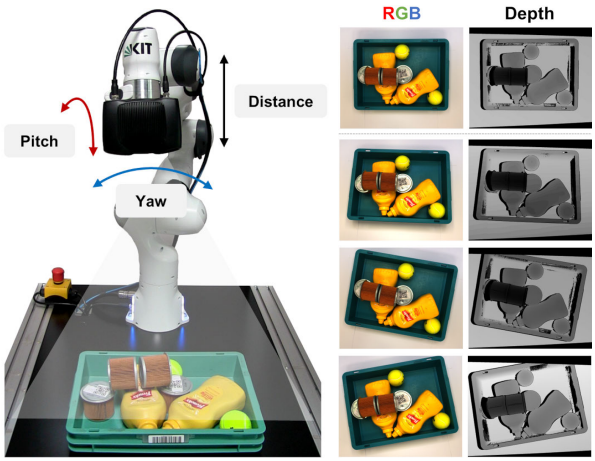


Fig. 5. (a) Camera poses are randomly sampled. (b) Data is captured from four different camera viewpoints: top-down and three randomly sampled poses.

manipulation order by traversing the graph coming from the target object vertex.

IV. Real-World Data

In addition to our synthetic data (cf. Sec. III), we contribute comprehensive and fully-annotated real-world data.

A. Data Recording

Multi-modal real-world data is captured with a high performance RGBD camera system (Zivid Two) mounted at the robot's end effector from four different viewpoints: one top-down view and three randomly sampled poses in the bin hemisphere. The camera distance to the bin ranges from 0.6 m to 0.68 m, the pitch and yaw angle between $[0, 0.08\pi]$ and $[-0.06\pi, +0.06\pi]$, respectively (cf. Fig. 5). The parameters have been chosen to be kinematically feasible for the robot arm and to ensure that the bin scene is in the camera's field of view. Extrinsic (and intrinsic) camera calibration parameters are provided for all data samples.

B. Annotation

Due to the complexity of our label types and the domain knowledge required to annotate bin scenes, we utilized a pool of in-house experts to perform all annotations. Annotators were trained and familiarized with the objects and the physical setup. This was particularly crucial for tasks like annotating grasp labels and creating object layout graphs, as these tasks are challenging to automate and heavily rely on the human expertise. For labeling segmentation masks, a semi-supervised approach [51] based on click-based interactive segmentation was used, allowing for efficient annotation of visible panoptic object masks. The occluded parts of objects (if any) are manually annotated through human expertise and assigned to the corresponding visible object masks based on their unique instance IDs. Together, they form panoptic amodal masks (cf. Fig. 6(e)). To accurately label the 6-DoF object poses, an annotation tool has been developed that combines the panoptic segmentation mask with the Open3D point-to-point ICP algorithm [52] to provide an initial estimate of an object's pose. The estimated pose is then validated and fine-tuned by a human annotator through a two-step process. First, the annotator manually adjusts the 3D model to roughly align with the object point cloud. Then, either the Open3D point-to-point or the point-to-plane ICP algorithm [52] can be used to further refine the pose. This approach not only ensures accurate pose estimation but also increases efficiency by leveraging automation and human expertise. However, due to occlusion, depth sensing, and object symmetry, labeling some scenes and objects can be challenging. To provide transparency in our dataset, a list of objects and scenes is provided for which labels may be less confident. Curating a comprehensive real-world dataset is time-intensive, especially if the considered label types require substantial human supervision, as in our case. The total time spent to annotate our real-world data is estimated to be 150 hours, not including the time for data collection, annotation tool development, and data processing.

V. Sim/ Real Datasets Details

MetaGraspNetV2 consists of two datasets: a large-scale synthetic training dataset MGN-Sim and a real-world evaluation and test dataset MGN-Real. MGN-Sim has 296k samples distributed over 8k scenes and 37 viewpoints, while MGN-Real contains 3.2k samples distributed over 800 scenes and 4 viewpoints.

A. Label Overview and Dataset Statistics

For both datasets a comprehensive collection of labels types is provided. In MetaGraspNetV2, we are able to provide the same extensive annotation types for MGN-Real as for MGN-Sim (see Fig. 6 and Table II). For MGN-Real, visible semantic and instance (panoptic) masks are available for all scenes and all viewpoints. Amodal panoptic segmentation masks, object relationship graphs, vacuum grasp regions, as well as parallel-jaw grasps are annotated for the first 500 scenes and top-down viewpoint (MGN-Real500). For MGN-Sim, all label types are available for all viewpoints. For an overview over label types and quantities for both synthetic and real dataset see Table II. Fig. 7 illustrates our dataset statistics regarding number of objects per scene and object size with a comparison to the WISDOM dataset [5]. In particular Fig. 7(b) shows that our designed MGN-Sim and MGN-Real datasets offer better coverage of relative object sizes compared to [5]. Note that the total number of instances in the WISDOM-REAL dataset is significantly lower (3849) than in MGN-Real (12122).

B. Task-Specific Difficulty Levels

Our dataset is designed to provide tailored configuration for different use cases and benchmark testing. We developed multiple levels of task-specific difficulties to facilitate research for vision-driven bin picking and allowing for customized performance monitoring of object and grasp detection methods.

TABLE II
LABEL COVERAGE IN METAGRASPNETV2

Task	Label Type	MGN-Sim	MGN-Real500
Grasping	Vacuum Grasps	✓	✓†
	Parallel Grasps	✓	✓
Object Detection	Visible Instance Masks	✓	✓†
	Amodal Instance Masks	✓	✓
	Visible Semantic Masks	✓	✓†
	Amodal Semantic Masks	✓	✓
Scene Reasoning	Object Layout Graph	✓	✓
	Occlusion Score	✓	✓
	6-DoF Object Poses	✓	✓

† label available beyond MGN-Real500 dataset

1) Grasp Point Detection: Challenging objects for robotic grasping are often categorized based on the physical properties affecting sensing and grasping [46], [53]. The most represented object categories are transparent (e.g. glass), highly reflective (e.g. metal sheets), and deformable objects (e.g. fabrics). When using vacuum technology, air-permeability and object surface become important factors as well. For MGN-Real, all object masks $O_i, i = 1...N$ are assigned ordinal scaled values 0-low, 1-medium, 2-high regarding the following properties: air-permeability $D_{i,\text{air}}$, object dimension w.r.t. gripper dimension $D_{i,\text{grip}}$, shape $D_{i,\text{shape}}$, specularity $D_{i,\text{spec}}$, texturedness $D_{i,\text{text}}$, occlusion $D_{i,\text{occl}}$ and transparency $D_{i,\text{trans}}$. While most attributes require only one label per object class, occlusion must be evaluated once per scene. The overall scene vacuum grasp difficulty score S_G is defined as the averaged sum over all property difficulties D_p and object masks (cf. Fig. 8(b)):

$$S_G = \sum_{p \in \text{Props}} \lambda_p \cdot \frac{1}{N} \sum_{i=1}^{N} D_{i,p}. \quad (1)$$

The weights λ_p control the impact of properties p based on the underlying use-case. For simplicity, all weights are assigned the same value $\lambda_p = 1$. 1D k-means clustering [54] is used to separate S_G into easy, medium and hard difficulty levels.

2) Object Detection and Semantic Segmentation: Scenes have a varying number of objects, which can add to the overall complexity of the scene. To capture the difficulty of the scene while also considering the difficulty of each individual object, we adopted a two-step approach. First, a per-object difficulty score is calculated, and then these scores are aggregated to obtain a scene difficulty score (cf. Fig. 8).

One preeminent challenge for object detection and segmentation is the heavy occlusion caused by the complex stacking relationship between objects, often confusing these systems to make false positive or false negative predictions. To better understand, we classify the scenarios leading to a false detection for an object O as (cf. Fig. 9):

- False Positive: Object O has multiple components due to occlusion.
- False Positive: Object O has holes.
- False Positive: The bounding box or amodal mask for object O may enclose components from other objects, which can lead to false detections in object detection and false segmentation in segmentation tasks.

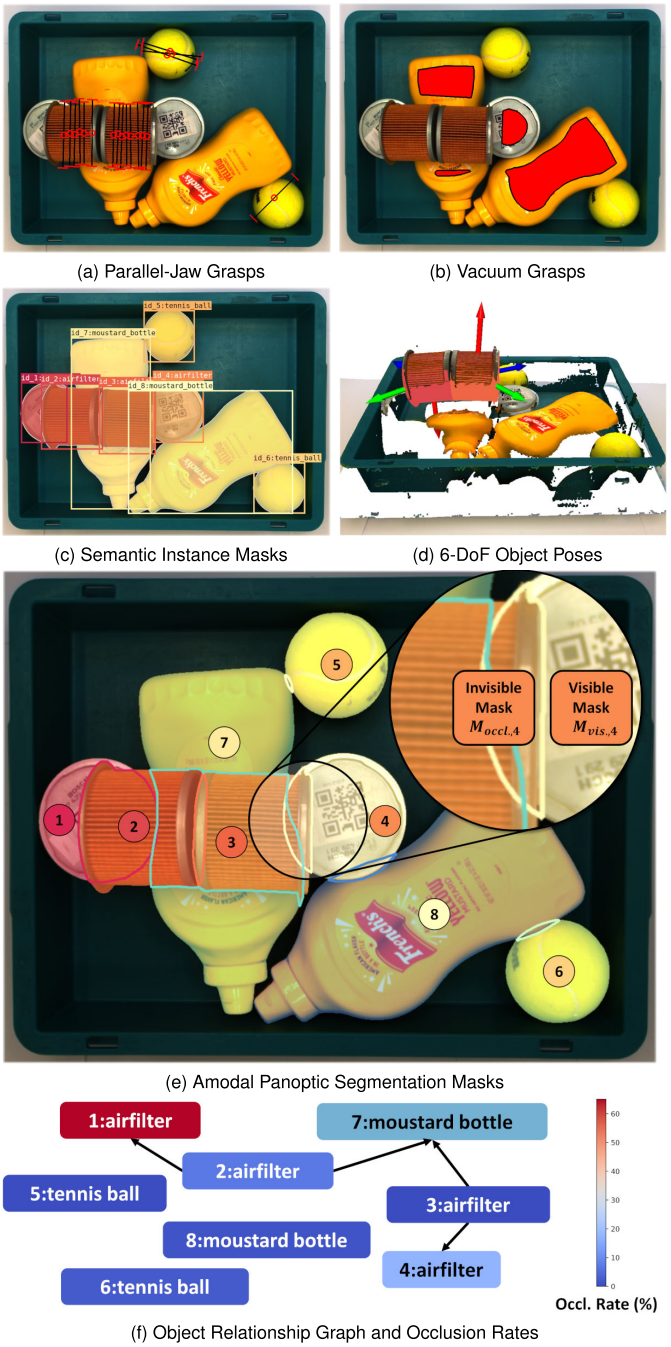


Fig. 6. Our real-world data provides annotations for top-down parallel-jaw and vacuum grasps, semantic instance segmentation masks, object poses, amodal panoptic segmentation masks, occlusion rates, and object relationship graphs.

- False Negative: The false detection is of the same class as O.
- False Negative: The false detection's mask is completely enclosed by object O's mask.

Still, not all false detections are equal. Objects with a larger mask within the bounding box are considered to be more important than those with a smaller mask. To address this issue, a weight term is proposed for each potential false detection. Let M_O be the object's mask and M_{F_j} be the mask of the jth potentially false detected object (cf. Fig. 8(a)).

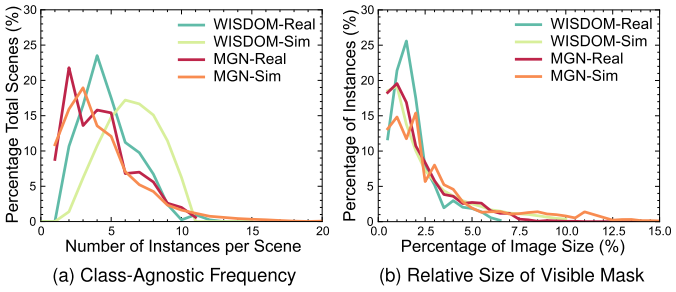


Fig. 7. MetaGraspNetV2 dataset characteristics with comparisons to WISDOM dataset [5]. As both datasets are designed for bin picking, many scenes contain multiple instances (a) with small sizes compared to image size (b).

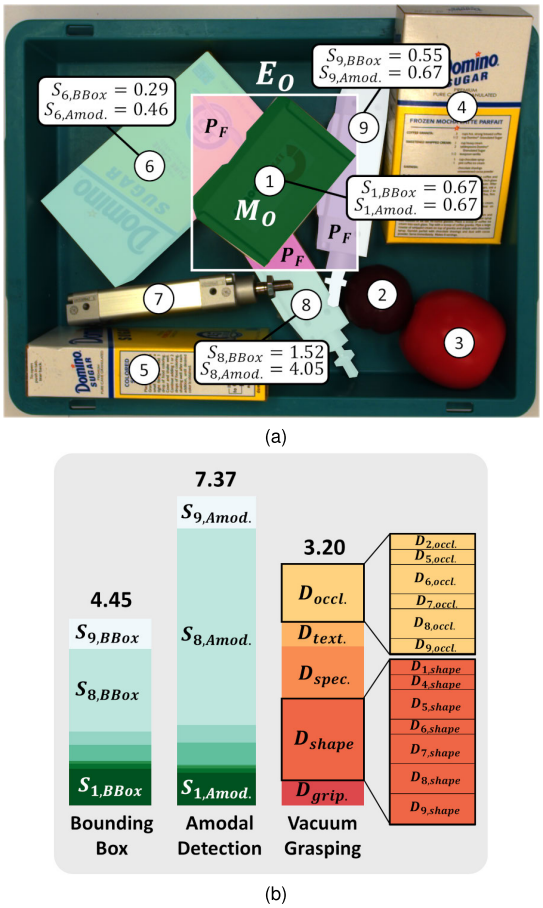


Fig. 8. Task-specific difficulty scores enable targeted performance evaluation. Scene-wise difficulty scores for object detection are calculated by adding object-wise difficulties for bounding box and amodal detection, as illustrated in (a) and (b) for object instances 1, 8 and 9. Scene-wise vacuum grasp difficulty levels are obtained by adding up property difficulties, as illustrated for D_{occl} and D_{shape} in (b). For visualization purposes, scaling alternates.

An enclosure mask E_{O} is defined in a task-specific manner, depending on the underlying task. For visible object detection, the enclosure mask E_{O} is defined as the bounding box of O's visible mask. For amodal object detection, the enclosure mask E_{O} is specified as the bounding box of O's amodal mask. The mask $P_{\text{F}_j} = E_{\text{O}} \cap M_{\text{F}_j}$ for the potential false detection object F_j is the intersection of the enclosure mask E_{O} and F_j's mask M_{F_j} (cf. Fig. 8(a)). The potential false detection is weighted by the ratio $R_{\text{F}_j} = \frac{|P_{\text{F}_j}|}{|M_{\text{O}}|}$ between the object mask size and the false detection mask size, where function $|.|$ computes the number

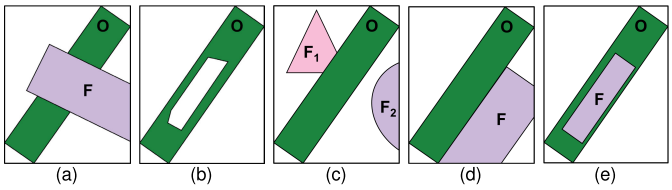


Fig. 9. Five false detection scenarios for an object O: (a) object F cross-cuts object O into two components, (b) object O has a hole, (c) the bounding box of object O contains components from objects F_1 and F_2, (d) object F has the same class as object O and significant overlap, and (e) object F is completely enclosed by object O.

of pixels in the mask. However, this simple ratio will lead to an unbounded value for the false detection weight when the object's mask M_{O} is small and the false detection mask P_{F_j} is large. We specify that the maximum weight of a false detection can only be 1, even though R_{F_j} might be larger. The difficulty score associated with jth potential false detection is defined as $S_{\text{c}_j} = \text{clip}(R_{\text{F}_j})$, where clip(.) is a function that clips the value to [0, 1]. When there are k potential false detections, the proposed object-wise difficulty score S_{O} is composed as the sum over all the weighted potential false detections:

$$S_{\text{O}} = \sum_{j=1}^{k} S_{\text{c}_j} = \sum_{j=1}^{k} clip\left(\frac{|E_{\text{O}} \cap M_{\text{F}_j}|}{|M_{\text{O}}|}\right) \tag{2}$$

Note that this design may cause some of the false detections to be counted multiple times. We reason that each false detection has a different reference object O making them all unique.

Scenes typically feature varying numbers of objects with different sizes. In order to aggregate object difficulty scores into a scene difficulty score, two critical questions have to be considered. The first question is whether size needs to be included to weigh the per-object difficulty score. To address this, we examine typical object detection networksMask R-CNN [55] and YOLO [56]. Both networks are capable of performing scale-invariant object detection. Therefore, size is omitted as a factor in the difficulty score. The second question is to find an appropriate method for aggregating object-level difficulty scores into a scene difficulty score. While using the mean value is one possible solution, it may not be optimal as some bins are more filled than others and the items' spatial distribution within the bin can also vary. We motivate that an increase in the number of objects within the bin can result in a higher incidence of false detections. In a scene with N items, to aggregate N item-level difficulties S_{O_i} into a scene-level difficulty score S_{S}, we employ a straightforward summation approach: $S_{\text{S}} = \sum_{i=1}^{N} S_{\text{O}_i}$. To make the difficulty levels more user-friendly, we cluster scene difficulties into four discrete levels: amodal-basic, amodal-easy, amodal-medium, and amodal-hard. The basic level comprises objects spread out in the bin and without any occlusion. The remaining scenes in the dataset are then classified into easy, medium, and hard categories using 1D k-means clustering [54] based on their difficulty scores.

C. Novel Objects List

Having accurate 3D scans for all objects is unrealistic for real-world applications, e.g. product range or packaging are

TABLE III
NOVEL OBJECTS AND PROPERTIES

Class	Non-Convex	Black	Varying Shape	Transparent
Pear	✓		✓	
Mug	✓			
Power drill	✓		✓	
Black clamp	✓	✓	✓	
Black marker		✓	✓	
Cable	✓		✓	
Cable in transparent bag			✓	✓
Wineglass	✓		✓	✓
Eyeglass	✓	✓	✓	✓
Crayon Box				
Greek Bust	✓		✓	
Object in plastic foil				✓
Object in bubble wrap				✓
Wooden Block			✓	
Mannequin	✓		✓	

likely to change. Therefore, it is crucial to test performance on novel objects and scenes to ensure functionality at application time. A novel object test set MGN-Novel is constructed based on the following properties (cf. Table III): convex, non-convex, transparency, varying shape depending on orientation, and color. Non-convex objects are harder to detect and objects with varying shape can be hard to detect in their entirety. Transparency and color pose special challenges to the sensor system and can results in false or noisy measurements.

VI. OCCLUSION-AWARE FAST PICKING

MetaGraspNetV2's comprehensive label set allows the development of holistic vision systems. Understanding the occlusion properties of objects in the grasp scene can optimize robot capacity by allowing to pick multiple objects in a single image acquisition and processing step, effectively tackling the challenge of fast-cycle times for robotic picking systems (cf. Sec. I). Occlusion-aware fast picking comes without any additional costs and can be applied together with conventional approaches including physical cell layout optimization, trajectory planning, optimization of sequence control or algorithms.

A picking robot's vision sensor can be in-hand or externally mounted above the bin. In-hand camera systems allow for more flexibility regarding viewpoints and cell design, while external camera mounts reduce end effector loads and dimensions. For both designs, in order to acquire a new image, the robot has to stop working, either by moving to the desired camera viewpoint pose (in-hand mount) or by moving out of the scene (external mount). Considering that a new image is normally captured and processed for each grasp (single-shot-single-pick), this results in large idle times when multiple objects have to be picked. To address this issue, we propose a single-shot-multi-pick (SSMP) method (cf. Fig. 10) for picking multiple items based on a single camera shot. SSMP consists of three modules: A) detecting grasps in a given scene (cf. Sec. VI-A), B) localizing objects in the scene and reason about their occlusion properties (cf. Sec. VI-B), and C) combining both results for object detection and grasp prediction to deduce suitable multi-pick manipulation sequences (cf. Sec. VI-C). All modules can be trained exclusively on MGN-Sim, no external data sources are needed.

A. Grasp Detection

For a given depth image I of a bin scene, multiple grasp candidates G_k are predicted by a grasp detection network

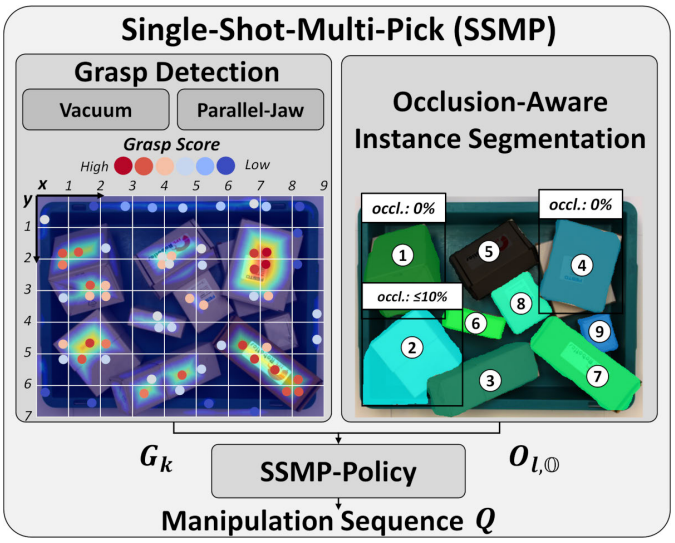


Fig. 10. Block diagram of single-shot-multi-pick (SSMP) method. Grasps G_k (here illustrated for vacuum grasping) and object detection masks together with occlusion properties $O_{I,\mathbb{O}}$ are connected in the SSMP-policy to reason about suitable multi-pick manipulation sequences Q.

$f^{\text{grasp}} = \{f^{\text{sc}}, f^{\text{pj}}\}$. A grasp candidate G_k is represented by its center location $(x, y)_k$ in 2D pixel space and a grasp quality score s_k, as illustrated in Fig. 10 and explained in more detail for vacuum and parallel-jaw grasping in this section:

$$G_k = \{(x, y), s\}_k. \tag{3}$$

1) Vacuum Grasp Detection (MGN-SG): In line with recent work [2], [38], we interpret vacuum grasp detection as a pixel-wise graspability learning problem and propose a vacuum grasp detection method MGN-SG. Given a depth image I, a vacuum grasp heatmap V is predicted by a deep vacuum grasp segmentation network f^{sc} based on DeepLabv3 [57] architecture. Contrary to [2] and [38], interpreting vacuum grasp prediction as a pixel-wise regression task, we define it as a pixel-wise classification task over $C = 25$ bins $C_i = i/(C-1), i = 0 \dots C-1$ [58], representing grasp score values in the range of $[0, 1]$ with a resolution of 0.04. Cross-entropy loss $\mathcal{L}$ is used to train the network's weights θ^{sc}:

$$\mathcal{L}(y, \theta, I) = \frac{-1}{HW} \sum_{x,y=0}^{H,W} \sum_{c=0}^{C-1} s^{\text{gt}}_{x,y,c} \cdot \log(f^{\text{sc}}_{x,y,c}(I, \theta^{\text{sc}})), \tag{4}$$

where H, W is the height and the width of the image and s^{gt} represents ground truth. At prediction time, the pixel-wise vacuum grasp heatmap V is obtained by: $V = \text{softmax}(f^{\text{sc}}(I))$.

In order to find reliable vacuum grasps, it is desirable to select grasp points located at the center of vacuum grasp areas, far away from object edges or boundaries. Related work, such as [2] merges object center and vacuum grasp prediction. However, we found that, especially for complex, non-convex items or cluttered scenes, a simple convolution operation of both heatmaps often harms the quality of predicted grasps and frequently results in failed grasp attempts. In our work, we focus on detecting reliable vacuum seal. We propose a three-step post-processing approach that rewards

grasp points in V when they are located in the center of predicted vacuum grasp areas. In the first step, based on a threshold $\lambda^{\text{sc}}=0.7$, V is segmented into j high-scoring vacuum regions $V_j | V(x,y)_j > \lambda^{\text{sc}}$ applying `regionsprops` function in scikit-image library [59]. In the next step, each pixel (x, y) in V is assigned a distance value greater or equal ϵ with regard to the distance to its region's boundary ∂V_j. The resulting distance heatmap D follows to:

$$D(x,y)_j = \begin{cases} \max(\text{d}((x,y), \partial V_j), \epsilon) & \text{if } (x,y) \in V_j \\ \epsilon & \text{if } (x,y) \notin V_j. \end{cases} \quad (5)$$

The distance function d(.) is implemented as `medial_axis` method in scikit-image [59]. The constant $\epsilon \ll 1$ is added in case no region V_j is found. In the last step, the two resulting heatmaps V and D are combined via 3D convolution into one final prediction heatmap $H = \text{Conv3D}_{(2\times2)}(V, D)$ (cf. Fig. 10). Following the grid-sample approach from [2], $k=1\ldots K$, $K = 512$ grasps $G_k = \{(x, y), s\}_k$ are sampled per scene by rasterizing the prediction heatmap H into 10×10 pixel large sub-grids $\tilde{H}$ and searching for their local maxima:

$$(x,y)_k = \arg\max_{\tilde{x},\tilde{y}}(\tilde{H}). \quad (6)$$

The local maxima, as illustrated in Fig. 10 with colored dots, correspond to the predicted grasp quality scores s_k from (3).

2) Parallel-Jaw Grasp Detection (MGN-PJ): As proposed in [18] and widely applied, e.g. in [11], [29], and [60], we interpret parallel-jaw grasp detection as oriented bounding box (*grasp rectangle*) detection problem and propose a straightforward parallel-jaw grasp detection method MGN-PJ based on Oriented R-CNN architecture [61]. To cover the additional three degrees of freedom for describing top-down parallel-jaw grasp configurations [18], we extend our previous general grasp representation in (3) with an angle $\theta \in (-\pi/2, \pi/2]$ describing the gripper orientation relative to the vertical camera axis, a gripper opening distance d, and a finger width w. Our grasp prediction model f^{pj} predicts $k = 1\ldots K, K = 1024$ 2D grasps candidates G_k together with their associated grasp quality score s_k based on a depth image I.

B. Occlusion-Aware Instance Segmentation

In VMRD [6], layout detection is based on two consecutive stages: object detection and object pair relationship reasoning due to shared feature maps. We simplify the problem and focus on distinguishing between first-layer and second-layer objects, rather than predicting the relationship between *all* objects. First-layer objects are those that can be grasped without changing the pose of any other object in the scene, while the remaining objects are referred to as second-layer objects. Non-occluded objects are first-layer objects, occluded objects are associated to be second-layer objects. However, the appearance of objects with low occlusion values can be very different from heavily occluded objects. For example, an object can be occluded by an edge of another object, or simple touching, and still represents a second-layer object (cf. object pair (8, 4) in Fig. 10). To account for these characteristics, we introduce

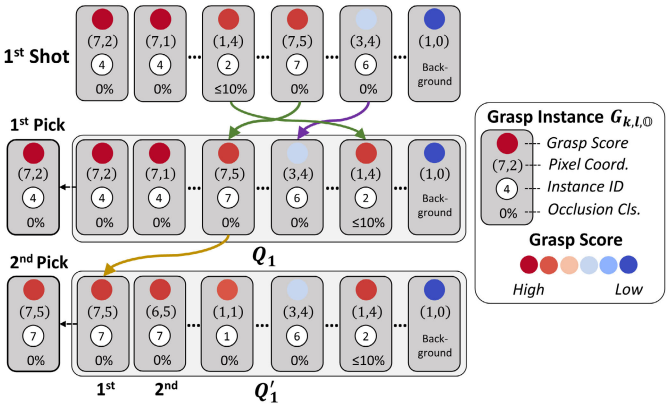


Fig. 11. Example for single-shot-multi-pick (SSMP) policy. **1st Shot**: For the proposed single-shot-multi-pick (SSMP) policy, grasps and object instances are linked with each other based on their coordinates. **Gray boxes** represent connected grasp instances, encapsulating grasp score, pixel coordinates, object id, and occlusion property. For visualization purposes, pixel coordinates are simplified to grid coordinates from Fig. 10. **1st Pick**: Grasps instances are ordered in a manipulation sequence queue Q based on a rule set that considers occlusion properties and grasp scores. The robot executes the first entry in Q_1 and grasps item 4. **2nd Pick**: Q_1 is updated by removing all entries associated to item 4. A second pick is executed grasping item 7. SSMP initiates picks as long as the system is sure that the scene layout has not changed unexpectedly. If not, it restarts the process by acquiring a new image.

a second occlusion class for objects with less than 10% occlusion. Our proposed occlusion-aware instance segmentation network f^{obj} is based on straightforward Mask R-CNN [55] architecture. Given a depth image I, object masks O_l are predicted together with their occlusion class $\mathbb{O}$ (see Fig. 10). Ground truth is obtained by assigning visible instance mask labels $M_{\text{vis},l}$ the following occlusion classes $\mathbb{O}$ based on their amodal masks: unoccluded, less than 10%, and greater than 10%:

$$\mathbb{O} = \begin{cases} \mathbb{O}_{0\%} & \text{if } M_{\text{occl},l} \,/\, (M_{\text{vis},l} + M_{\text{occl},l}) = 0 \\ \mathbb{O}_{\leq 10\%} & \text{if } M_{\text{occl},l} \,/\, (M_{\text{vis},l} + M_{\text{occl},l}) \leq 0.1 \\ \mathbb{O}_{>10\%} & \text{if } M_{\text{occl},l} \,/\, (M_{\text{vis},l} + M_{\text{occl},l}) > 0.1. \end{cases} \quad (7)$$

More details on the multi-task loss for Mask-RCNN training can be found in [55].

C. Policy

Our proposed single-shot-multi-pick (SSMP) policy reduces the idle time of a picking robot, which is otherwise used for image acquisition and processing. The goal is to perform multi-pick sequences with only *one* image acquisition without sacrificing grasp reliability.

For an initial grasp scene with N objects, the algorithm (cf. Algorithm 1) starts by capturing a depth image I_i of the scene at time step i. Given f^{grasp} (cf. Sec. VI-A) and f^{obj} (cf. Sec. VI-B), grasps $G_{i,k}$ and object masks with their occlusion classes $O_{i,l,\mathbb{O}}$ are predicted in the first step (cf. Fig. 10). In our notation, we use k to index predicted grasps and l to index detected object instances. Sharing the same 2D pixel-space, grasps $G_{i,k}$ can be assigned to objects $O_{i,l,\mathbb{O}}$ when possible (cf. *upper row* in Fig. 11): $G_{i,k,l,\mathbb{O}}$. Based on the predicted grasp quality score and assigned occlusion class, grasps $G_{i,k,l,\mathbb{O}}$ are sorted in a manipulation sequence queue Q_i according to the following rule set (cf. Fig. 11):

Algorithm 1 Single-Shot-Multi-Pick Algorithm

Parameters : N (number of objects to clear)
Input : f^{grasp}, f^{obj}
Output : L_{cleared}

$L_{\text{cleared}} \leftarrow \{\}$
while len(L_{cleared}) $< N$ **do**
  $blind = 0$
  capture image I_i at time t_i
  detect $G_i = f^{\text{grasp}}(I_i)$ and $O_i = f^{\text{obj}}(I_i)$
  create manipulation queue $Q_i(G_i, O_i)$
  while len(L_{cleared}) $< N$ **do**
    $Q_i' \leftarrow$ filter Q_i for L_{cleared}
    $G_i^{\text{1st}} \leftarrow$ get head of Q_i'
    transform G_i^{1st} into world and execute grasp
    if *grasp successful* **then**
      if *object unoccluded* **then**
        $L_{\text{cleared}} \leftarrow \{O_{i,l,\mathbb{O}})$
        increment $blind$
      else
        $L_{\text{cleared}} \leftarrow \{O_{i,l,\mathbb{O}})$
        break
      end
    else
      break
    end
  end
end

1) Non-occluded grasps are prioritized over grasps belonging to occluded objects or background (cf. *green arrows* in Fig. 11). *Reason*: The system prioritizes unoccluded (first-layer) objects because they can be picked without changing the pose of the remaining objects in the scene.
2) Within a set of grasps belonging to the same occlusion class, grasps are ordered by their grasp score (cf. *purple arrow* in Fig. 11). *Reason*: Under consideration of 1), always select the grasp candidate with the highest score.

Once the manipulation queue $Q_i = \{G_{i,k,l,\mathbb{O}}^{\text{1st}}, G_{i,k,l,\mathbb{O}}^{\text{2nd}}, \dots\}$ has been created, the algorithm initiates the execution of the grasp at the beginning of queue Q_i (cf. *middle row* in Fig. 11). Grasps which are in collision with the scene are filtered out. If the pick has been successful, its underlying instance $O_{i,l,\mathbb{O}}$ is added to the list of cleared objects L_{cleared} and all remaining grasps in Q_i belonging to the same object O with index l are removed, resulting in Q_i'. For the next pick, the remaining grasp instances in Q_i' move up (cf. *yellow arrow* in Fig. 11) and the grasp at the beginning is executed (cf. *lower row* in Fig. 11). This process of updating Q is repeated for succeeding picks as long as the system is sure that the grasp scene has not changed unexpectedly. Therefore, if a grasp fails or an item has been picked which might have been occluded (either no object detection was available or occlusion was predicted), the robot exits the current control loop and acquires a new picture at time step $i + 1$ to restart the process. Interested readers are encouraged to view our supplementary video featuring SSMP robot picking demos for a vacuum and a pararallel-jaw gripper.

VII. Experiments

In our previous work [17], the potential of our synthetic data for vacuum grasping and class-agnostic instance segmentation has been demonstrated. In addition, the proposed vacuum seal model was able to generalize to different cup materials and dimensions with an average seal precision of 95.0% and outperformed state-of-the-art models, e.g. [1] and [2], in real-world experiments. The experiments in this work are designed to evaluate the potential of our data and methods with focus on dexterous grasping, robust object detection and fast cycle times, three key-challenges of robot picking systems (cf. Sec. I). For the task of vacuum and parallel-jaw grasping, the generalization capabilities of our data and algorithms to truly *unseen* real-world environments are evaluated. In addition, we extend our physical robot bin picking experiments with a thorough evaluation of the proposed SSMP method for occlusion-aware fast picking. Besides real robot test series, extensive experiments for object detection, amodal segmentation and pose estimation are conducted.

A. Vacuum Grasp Experiments

In our previous work [17], the effectiveness of our data generation pipeline has been evaluated by comparing the grasp performance of the vacuum grasp network SuctionNet-1Billion [2] trained on their proposed large-scale real-world dataset versus a version of SuctionNet-1Billion trained on our synthetic data. The results showed clearly that our synthetic data improves picking performance for unseen objects. However, the robotic cell used to collect the evaluation data to monitor training performance and the test environment were identical. In this work, the generalization towards unseen environments and objects is evaluated. To prevent data leakage and increase the meaningfulness of our results with regard to novel environments and object sets, a completely new picking cell is built for our real robot tests (cf. Fig. 12). The new cell design differs from the setup used for recording MGN-Real in various aspects: camera mounting position and orientation, light conditions, two bins instead of one bin with different dimension and colors, as well as unseen objects and new arrangements in the scene. In order to exploit the full potential of our two datasets and mimic a realistic deployment phase, training is been conducted on synthetic data MGN-Sim, while MGN-Real is used exclusively to monitor performance during the training process. Testing is performed in an unseen robotic environment.

The proposed vacuum grasp model MGN-SG (cf. Sec. VI-A) and SuctionNet-1Billion [2] are trained on MGN-Sim. A batch size of 16, Adam Optimizer, and a learning rate of 10^{-4} are used. Training samples contain viewpoints $\{8, 10, 12, 14\}$ and the experiment is terminated after 50k training batch iterations. Training is conducted on a single NVIDIA Quadro RTX6000 GPU and took about 27 h per model. Successful grasp rate is used as evaluation metric and model performance is tracked on grasp difficulty splits of MGN-Real (cf. Sec. V-B).

As shown in Fig. 13, [2] achieves the highest successful grasp rate after 30k training iterations with 69.7%, while

Fig. 12. Physical robot experiments are designed to mimic a real-world logistics use-case, where a robot is tasked with transferring items from one bin to another, as is the case in many order-picking processes. The camera system Zivid Two (not shown) is mounted above the bin, vacuum is generated by an air-powered venturi vacuum pump (Festo VN-07-M-I3-PQ2-VQ2-A). The suction cup (Festo ESS-20-CS) has 20 mm diameter. Test objects are divided in two difficulty levels: moderate split-1 and challenging split-2.

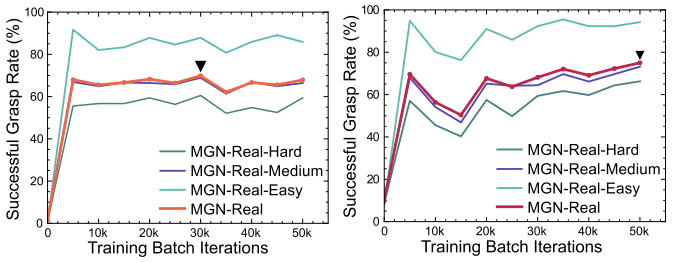


Fig. 13. Successful vacuum grasp rate evaluated during training for each 5k training iterations of SuctionNet-1Billion [2] (a) and MGN-SG (b). MGN-Sim is used for training, performance is monitored over all three difficulty levels of MGN-Real. Black arrows indicate the best overall result.

MGN-SG performs best after 50k iterations with 74.8%. For both networks, clear performance gaps between the difficulty splits in MGN-Real demonstrate that the proposed vacuum grasp difficulty rating is able to cover real-world challenges.

Physical Robot Experiment: For the robot experiments that follow a conventional single-shot-single-grasp (SSSP) policy, the following questions are of interest: How well does the proposed vacuum grasp model MGN-SG trained on MGN-Sim perform against state-of-the art networks [2], [62]? How large is the generalization error regarding novel objects and novel scenes comparing MGN-Sim against a large-scale real world dataset [2]?

Three different network architectures are considered: MGN-SG, FC-DexNet 4.0 [62], and SuctionNet-1Billion [2]. While emptying the bin, scene layout changes continually. Therefore, we focus in our real-world experiments on object level difficulties and divide our novel object test set in two levels: moderate split-1 and challenging split-2. Items are assigned

TABLE IV

REAL-WORLD VACUUM GRASP PERFORMANCE WITH STANDARD DEVIATION IN PARENTHESES FOR NOVEL OBJECTS IN CLUTTER AND UNSEEN ENVIRONMENTS

Policy	Method	Train Dataset	R_{grasp} (%)		R_{obj} (%)	
			Item Split-1	Item Split-2	Item Split-1	Item Split-2
SSSP	MGN-SG	MGN-Sim	87.6 (12.0)	74.6 (15.0)	94.0 (6.6)	87.5 (10.9)
	[62]	[62]	81.8 (14.7)	61.9 (16.2)	90.0 (8.4)	77.5 (12.2)
	[2]	MGN-Sim	69.8 (12.9)	55.1 (18.1)	84.0 (9.7)	73.0 (16.5)
	[2]	[2]	57.1 (15.7)	32.7 (10.6)	73.5 (14.2)	51.0 (14.5)
SSMP	MGN-SG	MGN-Sim	87.9 (10.2)	74.3 (14.7)	95.5 (5.9)	88.0 (10.3)
	[62]	[62] + MGN-Sim	86.8 (12.7)	70.0 (12.5)	93.5 (7.3)	85.5 (9.7)

to each split based on the following criteria (cf. Fig. 12): convex/non-convex shape, material, coverage of suctionable areas, specularity, and weight. Object arrangement in scenes is restored for all experiments across different methods. To align with experimental designs in previous works [2], [62], background is filtered out based on an empty bin depth image. A grasp is considered successful if the object is lifted up and stays in contact while moving to a predefined post-grasp pose. If the robot attempts to grasp the bin, the grasp is considered unsuccessful and assigned to the nearest object. After two failed grasp attempts per object and scene, a human supervisor removes the object. The weights for MGN-SG + MGN-Sim and [2] + MGN-Sim for the best training iteration (cf. Fig. 13) are selected and tested against provided pretrained versions of FC-DexNet 4.0[1] trained on synthetic data and SuctionNet-1Billion[2] trained on real-world data. Note that our test environment is completely different from the environment in which the evaluation data was collected. This allows for a fair comparison between methods. The number of successfully cleared objects over the number of total grasps attempts R_{grasp} (successful grasp rate) and the number of successfully cleared objects over the total number of objects R_{obj} (autonomously cleared object rate) are used as metrics. For each of the four test series, 400 items distributed over 40 bin scenes (20 runs of 10 split-1 objects and 20 runs of 10 split-2 objects) had to be picked, adding up to a sample size of 2040 grasp attempts. Reported numbers in Table IV and Fig. 14 are averaged across all runs and reported with standard deviation.

Table IV shows that the proposed MGN-SG method outperforms all other networks in terms of successful grasp rate R_{grasp} and autonomously cleared object rate R_{obj} for easy as well as challenging items. Depth image based methods (MGN-SG and [62]) perform better than RGBD-based method [2]. Comparing the grasp performance for [2] trained on MGN-Sim with the provided version trained on real-world data, it can be shown that our proposed synthetic data outperforms real-world data for R_{grasp} and R_{obj} by a large margin (cf. Table IV and Fig. 14(a)-(b)). This result confirms our previous experiments on novel objects [17] and extends it to novel environments.

For evaluating our proposed **single-shot-multi-pick (SSMP)** method, the following questions are of interest: How effective is SSMP compared to a conventional (SSSP) policy in terms of required image acquisitions? Is there a trade-off

[1]https://berkeleyautomation.github.io/gqcnn/index.html
[2]https://github.com/graspnet/suctionnet-baseline

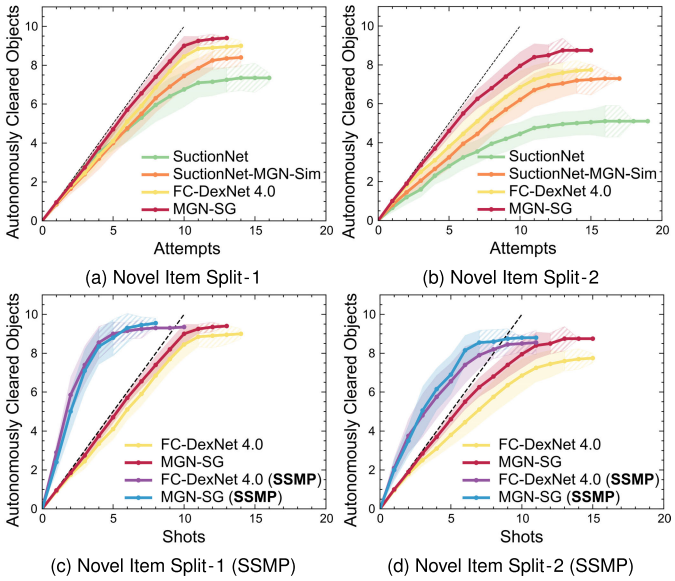


Fig. 14. Results from physical robot **vacuum** grasp experiments in clutter. In (a) and (b), the number of autonomously cleared objects is plotted over the total number of grasp attempts. The dashed line represents the ideal case, where all items are cleared with the first trial. In (c) and (d), the number of autonomously cleared objects is plotted over the number of image acquisitions for the proposed single-shot-multi-pick (SSMP) method. Top left means better performance, characterized by a high number of autonomously picked objects and a low number of required image acquisitions. The dashed line represents a baseline for an ideal single-shot-single-pick (SSSP) policy, where every grasp attempt is successful, but still requires an new recording. Standard deviation is scaled by a factor of 0.5 and shown in shaded and hatched areas for ≥ 10 active runs and < 10 active runs, respectively.

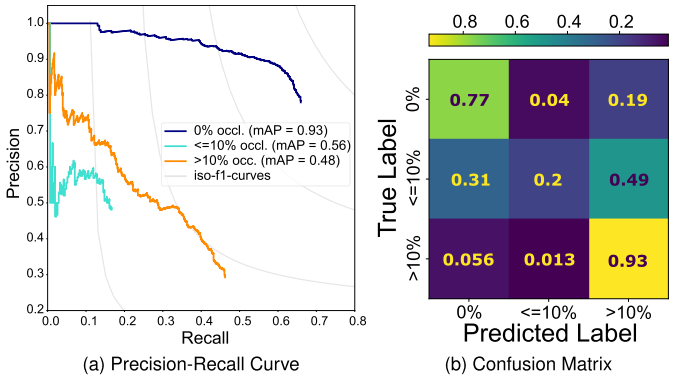


Fig. 15. Occlusion-aware instance segmentation network f^{obj} evaluated on MGN-Real500 with a confidence threshold of 0.7. A prediction is considered true positive if the correct occlusion class is predicted, $IoU \geq 75\%$, and no other prediction has been assigned to the ground truth label yet. The precision-recall plot in (a) shows that especially unoccluded items are reliably detected with more than 80% precision up to a recall value of 60%.

Fig. 16. (a) Robot setup for parallel-jaw grasp experiments. The electrical two-finger parallel gripper (Franka Hand) was customized with enlarged 3D printed fingers to avoid collision with the bin while picking objects close to the wall. (b) Test items are split in two sets: a subset of objects from previous vacuum grasp experiments (split-3) and parallel-jaw only objects (split-4).

between blind picking and grasp reliability? How well does the SSMP policy work together with a state-of-the-art vacuum grasp detection network [62]?

We have extended FC-DexNet 4.0 [62] with our SSMP policy. The proposed occlusion-aware instance segmentation network f^{obj}, a key component of SSMP, is trained on MGN-Sim using the same parameters as for training MGN-SG. For comparability, identical object scenes from previous experiments are recreated. As shown in Fig. 14(c)-(d), the number of cleared objects for our SSMP policy converges significantly faster than for the conventional SSSP policy. This observation is valid for both grasp detection methods MGN-SG and FC-DexNet 4.0 [62]. On average, to clear a scene of 10 split-1 objects, SSMP policy requires only 5.0 (σ=1.5) and 5.2 (σ=1.9) shots for MGN-SG and [62] method, respectively, reducing image acquisitions by 53.9% and 53.6% compared to the SSSP policy. For split-2 objects, the required image acquisitions for MGN-SG and [62] are reduced by 41.0% and 46%, respectively. Note that this increase in speed does not alter the grasp performance. In fact, Table IV shows clearly that SSMP policy even improves R_{grasp} and R_{obj}. Especially FC-DexNet 4.0 [62] seems to be profiting the most.

To gain further insight into the capabilities of SSMP, additional offline experiments on MGN-Real500 are conducted,evaluating f^{obj} for the task of occlusion-aware instance segmentation (cf. Fig. 15). For the reliability of SSMP policy, a high precision for detecting unoccluded items is required. The results in Fig. 15(a) confirm this precondition with high precision-recall values ($\text{mAP}^{75}_{\mathbb{O}_{0\%}} = 0.93$) for unoccluded items (*blue curve*) and low confusion values for highly occluded objects in Fig. 15(b). The observed overall low recall values for occluded objects (*turquoise* and *orange* in Fig. 15(a) do not have a large negative influence on the SSMP policy, since the algorithm focuses on detecting unoccluded objects.

B. Parallel-Jaw Grasp Experiments

When evaluating our data for the task of parallel-jaw grasping, the following questions are of interest: How well does a parallel-jaw grasping model trained on our synthetic data perform against state-of-the-art [28] in the real-world? How effective is our SSMP method for parallel-jaw grasping?

Physical Robot Experiment: Analogous to the previous vacuum grasp experiments, the real-world test environment is unseen and a scene consists of 10 objects arranged in high clutter (cf. Fig. 16(a)). Test objects are novel and divided in

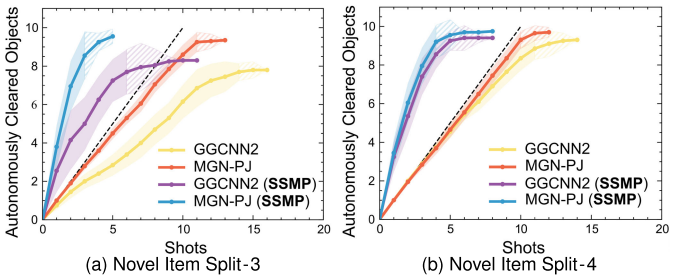


Fig. 17. Results from physical robot **parallel-jaw** grasp experiments in clutter. The number of autonomously cleared objects is plotted over the number of image acquisitions for the proposed single-shot-multi-pick (SSMP) method (*purple* and *blue*) and a conventional single-shot-single-pick (SSSP) policy (*yellow* and *red*). Top left means better performance, characterized by a high number of autonomously picked objects and a low number of required image acquisitions. The dashed line represents a baseline for an ideal single-shot-single-pick (SSSP) policy, where every grasp attempt is successful, but still requires an new recording. Standard deviation is scaled by a factor of 0.5 and shown in shaded and hatched areas for $\geq$ 10 active runs and $<$ 10 active runs, respectively.

TABLE V
REAL-WORLD PARALLEL-JAW GRASP PERFORMANCE WITH STANDARD DEVIATION IN PARENTHESES FOR NOVEL OBJECTS IN CLUTTER AND UNSEEN ENVIRONMENTS

Policy	Method	Train Dataset	R_{grasp} (%)		R_{obj} (%)	
			Item Split-3	Item Split-4	Item Split-3	Item Split-4
SSSP	MGN-PJ	MGN-Sim	84.4 (12.0)	92.5 (9.9)	93.5 (7.3)	97.0 (5.6)
	[28]	[19]	60.6 (18.9)	81.6 (12.7)	78.0 (15.4)	93.0 (7.8)
SSMP	MGN-PJ	MGN-Sim	84.5 (10.9)	86.5 (10.6)	95.5 (6.7)	97.5 (5.4)
	[28]	[19] + MGN-Sim	65.5 (14.2)	78.1 (11.0)	83.0 (9.0)	94.0 (8.6)

two separate item splits (cf. Fig. 16(b)): a mixed item split-3, equally balanced between split-1 and split-2 objects, and a parallel-jaw only split-4 containing objects which can only be grasped by a finger gripper due to weight, material, or shape properties. Like in the previous experiments, each test series consists of 400 items spread across 40 different bin scenes (20 runs of split-3 objects and 20 runs of split-4 objects), adding up to a sample size of 1893 grasp attempts. The same metrics R_{grasp} (successful grasp rate) and R_{obj} (autonomously cleared object rate) are used for evaluation.

We train MGN-PJ (cf. Sec. VI-A) for 12 full epochs on our synthetic MGN-Sim dataset. A batch size of 2, SGD optimizer, and a learning rate of 0.02 are used. Training samples contain 3 random viewpoints per scene. For performance comparison, the widely applied GG-CNN2 grasping network [28] is used. It is trained on synthetic depth images from large-scale Jacquard dataset [19] and is well established for grasping in clutter.

Looking at the experimental results for the conventional single-shot-single-pick (SSSP) policy in Table V, one can see that MGN-PJ trained on MGN-Sim outperforms GG-CNN2 [28] in terms of R_{grasp} and R_{obj} for both split-3 and split-4 objects by a large margin. Combining both item splits, MGN-PJ is able to autonomously pick 95.3% of objects (R_{obj}) with an average precision of 88.5% (R_{grasp}). These results show that our synthetic dataset allows the training of reliable parallel-jaw grasp detection methods in the real-world.

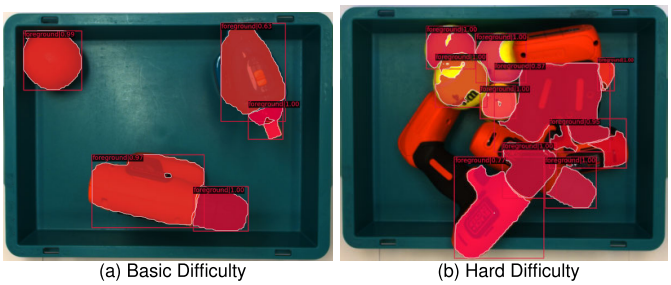
To evaluate the potential of the proposed **single-shot-multi-pick (SSMP)** method for parallel-jaw grasping, GG-CNN2 [28] is extended with our SSMP policy and the previous real-world grasp experiments are repeated. As shown in Fig. 17, with SSMP, the number of cleared objects converges significantly faster for both methods, MGN-PJ and GG-CNN2, as for the conventional SSSP policy (*blue* and *purple* vs. *orange* and *yellow* lines in Fig. 17). On average, for scenes containing objects from mixed item split-3, SSMP requires only 3.5 ($\sigma = 1.07$) and 5.7 ($\sigma = 2.19$) shots for MGN-PJ and [28] method, respectively, reducing needed image acquisitions by 68.2% and 57.5% compared to the conventional SSSP policy. For split-4 objects, image acquisitions are reduced to 4.1 ($\sigma = 1.34$) and 4.7 ($\sigma = 1.27$) shots for MGN-PJ and [28], resulting in 61.1% and 59.3% less image acquisitions. However, when comparing R_{grasp} and R_{obj} for both SSSP and SSMP policies (cf. Table V), this efficiency gain comes for split-4 objects with a small decrease in grasp precision (R_{grasp}). During our experiments, we noticed that in some rare cases, when grasping an object, the finger of the parallel-jaw gripper can unintendedly change the pose of background objects, resulting in possible failed downstream picks. This is particularly relevant for objects whose shape or pose might easily be changed (deformable textiles or round objects, cf. split-4 in Fig. 16(b)). Note that this was only observed for R_{grasp} and item split-4, considering autonomously cleared objects (R_{obj}), SSMP performs even better. Overall, the results confirm the effectiveness of our proposed SSMP policy for parallel-jaw grasping and extend our previous finding for the vacuum gripper in Sec. VII-A to ambidextrous grasping.

C. Object Detection Experiments

Object detection and segmentation are crucial but challenging tasks and are essential for achieving a comprehensive and dependable understanding of grasp scenes. These tasks enable the model to deliver consistent performance even in the presence of diverse object layouts, including defective or previously unseen objects. To evaluate the effectiveness of our dataset for these tasks, we designed two experiments. We first describe the setup that is shared by both experiments and then introduce the two experiments in detail.

1) Experimental Setup: To evaluate object detection, segmentation, and amodal segmentation tasks, the network architecture proposed in [14] UOAIS-Net was utilized. The architecture is implemented using a ResNet-50 [63] backbone. To facilitate comparison across datasets, all objects are treated as one class during training and evaluation. The learning rate is set to 0.015 and all models are trained for 12 iterations over the training set (epochs). The results are evaluated using the standard Microsoft Common Objects in Context (COCO) [64] object detection and instance segmentation metrics, including Bounding Box Mean Average Precision (Box mAP), Segmentation Mean Average Precision (Mask mAP), and Amodal Segmentation Mean Average Precision (AMask mAP).

2) Difficulty Levels Experiment: Evaluating a model's performance on objects with varying degrees of difficulty ensures its ability to accurately detect and segment objects across various scenarios. In this experiment, UOAIS-Net is trained on our MGN-Sim dataset, referred to as UOAIS-Net + MGN-Sim.



(a) Basic Difficulty (b) Hard Difficulty

Fig. 18. Example visible object detection results of UOAIS-Net + MGN-Sim.

TABLE VI
EXPERIMENT ON THE DIFFICULTY OF VISIBLE OBJECT DETECTION, INSTANCE SEGMENTATION AND AMODAL DETECTION

Difficulty	Box mAP (%)	Mask mAP (%)	AMask mAP (%)
amodal-basic	85.4	85.5	86.1
amodal-easy	70.7	71.9	72.1
amodal-medium	66.1	66.8	67.6
amodal-hard	58.1	55.5	58.2
Novel	47.7	48.0	48.5

During training, 4 viewpoints from MGN-Sim are randomly selected. After training, we fine-tune UOAIS-Net + MGN-Sim on two empty bin scenes from our MGN-Real dataset to account for the sim-to-real gap. Empty scenes contain no labels, but provide enough information for the models to learn new lighting conditions and backgrounds. The model is fine-tuned with a learning rate of 1 until the model produces no false positive proposals, as indicated by the region proposal loss converging to zero. Taking less than a minute to complete with an average of only 6 epochs, this process was remarkably fast. The fine-tuned UOAIS-Net + MGN-Sim model is evaluated on the proposed four levels of difficulties for amodal detection on MGN-Real500. The results, as shown in Table VI and Fig. 18, demonstrate that as the difficulty level increases, the performance of our model drops correspondingly. This outcome validates the effectiveness of our proposed difficulty level design and provides valuable insights into the capabilities and limitations of our model.

3) Object Detection Generalizability Experiment: The purpose of this experiment is to test the generalizability of our dataset by comparing models with the same architecture and training policy trained on different datasets. Evaluating the performance of these models on previously unseen real-world objects and datasets allows us to assess their ability to perform well in novel real-world scenarios. For this experiment, we train two UOAIS-Net models: one on our MGN-Sim dataset and the other one on the UOAIS-SIM synthetic dataset proposed in [14]. The models are referred to as UOAIS-Net + MGN-Sim and UOAIS-Net + UOAIS-SIM, respectively. The same fine-tuning process is applied as in the previous difficulty levels experiment. Both models are evaluated for the task of visible object detection, instance segmentation, and amodal detection on our MGN-Novel dataset (cf. Sec. V-C), which includes novel object classes not seen before. Furthermore, performance of both models is evaluated for the task of object detection and instance segmentation on the WISDOM-Real dataset [5] (cf. Fig. 7). Note that this dataset lacks amodal labels. The results, as shown in Table VII, indicate that the model UOAIS-Net + MGN-Sim trained on our synthetic data outperforms the model trained on the UOAIS-SIM dataset by at least 10% for each task when applied to our MGN-Real500 real-world dataset. When evaluating on our MGN-Novel dataset and the WISDOM-Real dataset, the UOAIS-Net + MGN-Sim model performs competitively. With empty bin fine-tuning, UOAIS-Net + MGN-Sim outperforms UOAIS-Net + UOAIS-SIM by 3.4% in average across the three tasks. With regard to experiments on WISDOM-Real dataset, we observe that UOAIS-Net + MGN-Sim demonstrates a 3.1% advantage in object detection and a 4.3% advantage in instance segmentation performance after fine-tuning. In short, all the experimental results support the conclusion that our MGN-Sim dataset is better suited for generalizing to real-world unseen objects and datasets.

TABLE VII
AMODAL OBJECT DETECTION AND SEGMENTATION RESULTS

Train Dataset	Test Dataset	Tuned?*	Box mAP (%)	Mask mAP (%)	AMask mAP (%)
UOAIS-SIM	MGN-Real500		55.7	55.6	55.9
MGN-Sim	MGN-Real500		66.5	65.8	66.5
UOAIS-SIM	MGN-Real500	✓	55.8	55.9	56.4
MGN-Sim	MGN-Real500	✓	68	68.2	69.1
UOAIS-SIM	MGN-Novel		42.2	37.6	37.7
MGN-Sim	MGN-Novel		42.1	39.4	41.2
UOAIS-SIM	MGN-Novel	✓	43.1	37.7	38.1
MGN-Sim	MGN-Novel	✓	44.9	41.3	43.2
UOAIS-SIM	WISDOM-Real		52.3	50.4	
MGN-Sim	WISDOM-Real		50.8	50.8	
UOAIS-SIM	WISDOM-Real	✓	52.8	51.0	
MGN-Sim	WISDOM-Real	✓	55.9	55.3	

∗ "Tuned?" indicates whether the model has been fine-tuned on the empty bin scenes

D. 6D-Pose Estimation Experiments

Object poses represent rich scene labels and are often used in industy due to the wide availability of CAD data. To validate our dataset with regard to the task of pose estimation, a base GDR-Net model [8] is trained on MGN-Sim and its performance is evaluated on MGN-Real500 using the ADD(-S) [65] metric. ADD(-s) measures the average distance deviated between the point clouds of the ground truth and the predicted objects. To provide a threshold-invariant evaluation score, AUC scores are calculated by varying the threshold to a maximum of both 10 cm and 2 cm, similar to [23]. Results for MGN-Real500 and a hold-out 15% test split of MGN-Sim are shown in Table VIII, separated in two occlusion test splits: less than 10% and greater than 10%. The presence of occlusion can be challenging for pose estimation. As seen in Table VII, the AUC performance dropped noticeably at occlusion levels greater than 10%. This is especially noticeable for an AUC threshold of 2 cm.

E. Discussion and Remarks to Practitioners

Extensive physical robot experiments in novel scenes clearly demonstrate the effectiveness of our proposed synthetic dataset MGN-Sim and vacuum grasp detection method MGN-SG

TABLE VIII
EXPERIMENT ON 6-DOF POSE ESTIMATION

Test Dataset	Occlusion	AUC @ 10 cm	AUC @ 2 cm
MGN-Sim	all	0.968	0.894
	$\leq 10\,\%$	0.979	0.926
	$> 10\,\%$	0.937	0.807
MGN-Real500	all	0.944	0.773
	$\leq 10\,\%$	0.944	0.781
	$> 10\,\%$	0.943	0.755

for vacuum grasping. Specifically, MGN-SG outperforms SuctionNet-1Billion [2] and FC-DexNet 4.0 [62] by a large margin in terms of R_{grasp} and R_{obj} for moderate as well as challenging items. For novel objects and scenes, SuctionNet-1Billion trained on MGN-Sim outperforms the same model trained on a large-scale real-world dataset [2]. We attribute the superior generalization capabilities of our synthetic data for novel objects and environments to a better dataset diversity in terms of object layout and dataset size. Due to the hybrid data collection approach in [2], requiring human annotated 6-DoF object poses for each scene, their dataset has only 190 scenes compared to MGN-Sim dataset featuring 8k scenes. When dealing with large and diverse article sets, not all objects might be graspable with a vacuum suction cup. Extensive real robot experiments confirm the effectiveness of our data for parallel-jaw grasping, significantly expanding the range of graspable objects (cf. item split-4 in Fig. 16(b)). The results show that MGN-PJ trained on MGN-Sim is able to reliably grasps objects even in high clutter, outperforming GG-CNN2 [28] by a large margin in terms of R_{obj} and R_{grasp}.

The main aim of our proposed SSMP policy is to reduce required image acquisitions without sacrificing grasp performance. Our bin picking experiments show that SSMP policy speeds up cycle rates by significantly reducing image acquisitions by more than 47% on average for vacuum grasping and 64.7% for parallel-jaw grasping, even for scenes with high clutter and challenging item sets. Based on these findings, SSMP or similar strategies could become an integral part of future time-critical picking systems.

Experiments on object detection validate the effectiveness of the proposed difficulty levels. With increasing difficulty, the model performance decreased accordingly providing insights into our model's capabilities and limitations. Experiments for MGN-Novel demonstrate the generalization capabilities of MGN-Sim for the task of object detection and amodal segmentation outperforming [14] on novel objects. Experiments on WISDOM-Real [5] dataset for visible detection show that models trained on MGN-Sim generalize well to novel environments. More generally, both results indicate that our MetaGraspNetV2 dataset is diverse and representative, allowing object detection models trained on it to generalize to previously unseen objects and novel environments.

VIII. CONCLUSION

In this paper, a universal bin picking dataset has been designed to facilitate grasping in unstructured bin environments. Our proposed MetaGraspNetV2 dataset consists of two datasets: (i) a large-scale synthetic dataset with 296k samples; (ii) a real-world test dataset with 3.2k samples. Both datasets provide a comprehensive label set for a wide variety of tasks including ambidextrous grasping, panoptic amodal segmentation, object relationship reasoning, and pose estimation. Novel task-specific difficulty levels are proposed for our real-world dataset targeting visible and amodal object detection, as well as vacuum grasping. They allow efficient evaluation of robotic vision systems and provide meaningful insights into their capacities and limitations. Extensive experiments for object detection, instance segmentation, amodal detection, parallel-jaw, and vacuum grasping have demonstrated the superior generalization capabilities of our dataset with regard to unseen objects and environments, outperforming state-of-the-art task-specific datasets. Our novel single-shot-multi-pick (SSMP) policy increased the time-efficiency of bin picking. Combining occlusion and scene reasoning with robust grasp detection, SSMP reduces cycle time by picking multiple objects in one single image acquisition. Physical robot experiments for unseen objects and environments demonstrated its effectiveness for vacuum and parallel-jaw grasping, reducing the number of required image acquisitions by more than 47% while outperforming state-of-the-art bin picking methods. Overall, with its high-quality data and comprehensive label set, the proposed MetaGraspNetV2 dataset has the potential to enhance the development of dexterous bin picking systems in terms of speed and reliability.

ACKNOWLEDGMENT

The authors would like to thank Tim Robin Winter, Yu Zhou, Tobias Rothmann, Bastian Neussell, Yu Li, and Vinzenz Rau for their help with the dataset and David Faessen for his help with the control of the gripper.

REFERENCES

[1] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, "Dex-Net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 5620–5627.

[2] H. Cao, H.-S. Fang, W. Liu, and C. Lu, "SuctionNet-1Billion: A large-scale benchmark for suction grasping," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 8718–8725, Oct. 2021.

[3] J. Mahler et al., "Learning ambidextrous robot grasping policies," *Sci. Robot.*, vol. 4, no. 26, Jan. 2019.

[4] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-GraspNet: Efficient 6-DoF grasp generation in cluttered scenes," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13438–13444.

[5] M. Danielczuk et al., "Segmenting unknown 3D objects from real depth images using mask R-CNN trained on synthetic data," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 7283–7290.

[6] H. Zhang, X. Lan, X. Zhou, Z. Tian, Y. Zhang, and N. Zheng, "Visual manipulation relationship network for autonomous robotics," in *Proc. IEEE-RAS 18th Int. Conf. Humanoid Robots (Humanoids)*, Nov. 2018, pp. 118–125.

[7] M. Ding, Y. Liu, C. Yang, and X. Lan, "Visual manipulation relationship detection based on gated graph neural network for robotic grasping," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 1404–1410.

[8] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16606–16616.

[9] H. Liu, Y. Wu, F. Sun, B. Fang, and D. Guo, "Weakly paired multimodal fusion for object recognition," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 2, pp. 784–795, Apr. 2018.

[10] Y. Feng et al., "Towards robust part-aware instance segmentation for industrial bin picking," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 405–411.

[11] A. Zeng et al., "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," *Int. J. Robot. Res.*, vol. 41, no. 7, pp. 690–705, Aug. 2019.

[12] A. S. Periyasamy, M. Schwarz, and S. Behnke, "SynPick: A dataset for dynamic bin picking scene understanding," in *Proc. IEEE 17th Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2021, pp. 488–493.

[13] H. Zhang et al., "REGRAD: A large-scale relational grasp dataset for safe and object-specific robotic grasping in clutter," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2929–2936, Apr. 2022.

[14] S. Back et al., "Unseen object amodal instance segmentation via hierarchical occlusion modeling," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 5085–5092.

[15] D. Kalashnikov et al., "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Proc. Conf. Rob. Learn. (CoRL)*, vol. 87, Oct. 2018, pp. 651–673.

[16] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *Int. J. Robot. Res.*, vol. 37, nos. 4–5, pp. 421–436, Apr. 2018.

[17] M. Gilles, Y. Chen, T. R. Winter, E. Z. Zeng, and A. Wong, "MetaGraspNet: A large-scale benchmark dataset for scene-aware ambidextrous bin picking via physics-based metaverse synthesis," in *Proc. IEEE 18th Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2022, pp. 220–227.

[18] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 3304–3311.

[19] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 3511–3516.

[20] A. X. Chang et al., "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.

[21] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "GraspNet-1Billion: A large-scale benchmark for general object grasping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11441–11450.

[22] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *Proc. Int. Conf. Adv. Robot. (ICAR)*, Jul. 2015, pp. 510–517.

[23] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Proc. Robot., Sci. Syst.* Pittsburgh, PA, USA: MIT Press, Jun. 2018, doi: 10.15607/RSS.2018.XIV.019.

[24] Z. Liu et al., "OCRTOC: A cloud-based competition and benchmark for robotic grasping and manipulation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 1, pp. 486–493, Jan. 2022.

[25] H. Zhang, J. Peeters, E. Demeester, and K. Kellens, "A CNN-based grasp planning method for random picking of unknown objects with a vacuum gripper," *J. Intell. Robot. Syst.*, vol. 103, no. 4, pp. 1–19, Dec. 2021.

[26] C. Eppner, A. Mousavian, and D. Fox, "ACRONYM: A large-scale grasp dataset based on simulation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 6222–6227.

[27] K. Kleeberger et al., "Transferring experience from simulation to the real world for precise pick-and-place tasks in highly cluttered scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 9681–9688.

[28] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *Int. J. Robot. Res.*, vol. 39, nos. 2–3, pp. 183–201, Mar. 2020.

[29] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," in *Proc. Robot., Sci. Syst.* Pittsburgh, PA, USA: MIT Press, Jun. 2018, doi: 10.15607/RSS.2018.XIV.021.

[30] M. Gou, H.-S. Fang, Z. Zhu, S. Xu, C. Wang, and C. Lu, "RGB matters: Learning 7-DoF grasp poses on monocular RGBD images," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13459–13466.

[31] P. Hopfgarten, J. Auberle, and B. Hein, "Grasp area detection of unknown objects based on deep semantic segmentation," in *Proc. IEEE 16th Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2020, pp. 804–809.

[32] B. Zhao, H. Zhang, X. Lan, H. Wang, Z. Tian, and N. Zheng, "REGNet: Region-based grasp network for end-to-end grasp detection in point clouds," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13474–13480.

[33] C. Eppner, M. M. Arsalan, and D. Fox, "A billion ways to grasps—An evaluation of grasp sampling schemes on a dense, physics-based grasp data set," in *Proc. Int. Symp. Robot. Res. (ISRR)*, 2019, pp. 890–905.

[34] A. Iriondo, E. Lazkano, and A. Ansuategi, "Affordance-based grasping point detection using graph convolutional networks for industrial bin-picking applications," *Sensors*, vol. 21, no. 3, p. 816, Jan. 2021.

[35] A. Bernardin, C. Duriez, and M. Marchal, "An interactive physically-based model for active suction phenomenon simulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 1466–1471.

[36] F. Gabriel, M. Fahning, J. Meiners, F. Dietrich, and K. Dröder, "Modeling of vacuum grippers for the design of energy efficient vacuum-based handling processes," *Prod. Eng.*, vol. 14, nos. 5–6, pp. 545–554, Dec. 2020.

[37] X. Provot, "Deformation constraints in a mass-spring model to describe rigid cloth behaviour," in *Proc. Graph. Interface Conf.*, 1995, pp. 147–154.

[38] P. Jiang et al., "Learning suction graspability considering grasp quality and robot reachability for bin-picking," *Frontiers Neurorobot.*, vol. 16, pp. 1–12, Mar. 2022.

[39] W. Liu et al., "Robotic picking in dense clutter via domain invariant learning from synthetic dense cluttered rendering," *Robot. Auto. Syst.*, vol. 147, Jan. 2022, Art. no. 103901.

[40] Y. Domae, H. Okuda, Y. Taguchi, K. Sumi, and T. Hirai, "Fast graspability evaluation on single depth maps for bin picking with general grippers," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 1997–2004.

[41] G. Zuo, J. Tong, H. Liu, W. Chen, and J. Li, "Graph-based visual manipulation relationship reasoning network for robotic grasping," *Frontiers Neurorobot.*, vol. 15, Aug. 2021, Art. no. 719731. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2021.719731

[42] X. Li et al., "A sim-to-real object recognition and localization framework for industrial robotic bin picking," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 3961–3968, Apr. 2022.

[43] L. Zeng, W. J. Lv, Z. K. Dong, and Y. J. Liu, "PPR-Net++: Accurate 6-D pose estimation in stacked scenarios," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 4, pp. 3139–3151, Oct. 2022.

[44] S. Tyree et al., "6-DoF pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 13081–13088.

[45] R. Kaskman, S. Zakharov, I. Shugurov, and S. Ilic, "HomebrewedDB: RGB-D dataset for 6D pose estimation of 3D objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2767–2776.

[46] Y. Sun, J. Falco, M. A. Roa, and B. Calli, "Research challenges and progress in robotic grasping and manipulation competitions," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 874–881, Apr. 2022.

[47] F. Stulp, E. Theodorou, J. Buchli, and S. Schaal, "Learning to grasp under uncertainty," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 5703–5708.

[48] R. Murray, Z. Li, and S. Sastry, *A Mathematical Introduction to Robotic Manipulation*. Boca Raton, FL, USA: CRC Press, 1994.

[49] V. Makoviychuk et al., "Isaac gym: High performance GPU-based physics simulation for robot learning," 2021, *arXiv:2108.10470*.

[50] NVIDIA Developer. (2019). *NVIDIA Isaac Sim*. [Online]. Available: https://developer.nvidia.com/isaac-sim

[51] K. Sofiiuk, I. A. Petrov, and A. Konushin, "Reviving iterative training with mask guidance for interactive segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 3141–3145.

[52] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," 2018, *arXiv:1801.09847*.

[53] H. Tian, K. Song, S. Li, S. Ma, J. Xu, and Y. Yan, "Data-driven robotic visual grasping detection for unknown objects: A problem-oriented review," *Expert Syst. Appl.*, vol. 211, Jan. 2023, Art. no. 118624.

[54] A. Grønlund, K. G. Larsen, A. Mathiasen, J. S. Nielsen, S. Schneider, and M. Song, "Fast exact k-means, k-medians and Bregman divergence clustering in 1D," 2017, *arXiv:1701.07204*.

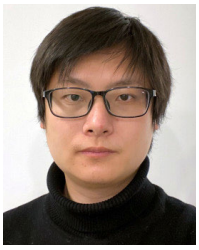
[55] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[56] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[57] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous convolution for semantic image segmentation," *arXiv:1706.05587*, 2017.

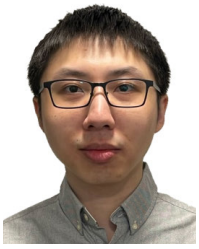
[58] A. N. Angelopoulos et al., "Image-to-image regression with distribution-free uncertainty quantification and applications in imaging," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 717–730.
[59] S. van der Walt et al., "Scikit-image: Image processing in Python," 2014, *arXiv:1407.6245*.
[60] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3355–3362, Oct. 2018.
[61] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3500–3509.
[62] V. Satish, J. Mahler, and K. Goldberg, "On-policy dataset synthesis for learning robot grasping policies using fully convolutional deep networks," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1357–1364, Apr. 2019.
[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[64] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," 2014, *arXiv:1405.0312*.
[65] S. Hinterstoisser et al., "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2012, pp. 548–562.

Maximilian Gilles received the B.Sc. and M.Sc. degrees in mechanical engineering from the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, in 2017 and 2020, respectively, where he is currently pursuing the Ph.D. degree. His research interests include robot perception and manipulation, with a special emphasis on material handling applications. He is also a member of the Robotics and Interactive System Group, Institute for Material Handling and Logistics (IFL).



Yuhao Chen (Member, IEEE) received the B.A.Sc. and Ph.D. degrees in electrical and computer engineering from Purdue University in 2015 and 2019, respectively. He is currently a Research Assistant Professor in systems design engineering with the Vision and Image Processing Laboratory (VIP), University of Waterloo. His research has been focused on developing computer vision and artificial intelligence solutions for industrial applications. He was a member of the Video and Image Processing (VIPER) Laboratory.

Emily Zhixuan Zeng (Member, IEEE) received the Bachelor of Applied Science degree in mechatronics engineering from the University of Waterloo in 2021, where she is currently pursuing the master's degree with the Vision and Image Processing Laboratory. Her research centers on computer vision, particularly addressing topics, such as object pose estimation and explainable AI with the University of Waterloo.



Yifan Wu received the B.Eng. degree in computing from the Imperial College London, London, U.K., in 2017, and the M.A.Sc. degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 2022. His research interests include optical remote sensing image processing, building extraction, and machine learning.

Kai Furmans (Member, IEEE) is currently a Professor in mechanical engineering and the Head of the Institute for Material Handling and Logistics, Karlsruhe Institute of Technology, Germany. His research interests include automation and robotics in material handling and modeling of such systems.

Alexander Wong is currently the Canada Research Chair of Artificial Intelligence and Medical Imaging, a member of the College of the Royal Society of Canada, a fellow of the Institute of Engineering and Technology and International Society for Design and Development in Education, the Co-Director of the Vision and Image Processing Research Group, and a Professor with the Department of Systems Design Engineering, University of Waterloo. He is also a P.Eng. He has published over 650 refereed journals and conference papers in various fields, such as computational imaging, artificial intelligence, computer vision, and multimedia systems.

Rania Rayyes received the Ph.D. degree in computer science (AI and robotics) from TU Braunschweig, Germany, in 2020. She is currently a Junior Professor with the AI & Robotics, Institute for Material Handling and Logistics, Karlsruhe Institute for Technology (KIT). Before joining KIT, she was a Post-Doctoral Researcher for two years with the Institute for Robotics and Process Informatics, TU Braunschweig. Her research interests include AI systems for real robot applications, human–robot leaning, and machine vision. She has received several awards during her academic career, including Robotics Talent Award for the Ph.D. dissertation.