



# MetaMVUC: Active Learning for Sample-Efficient Sim-to-Real Domain Adaptation in Robotic Grasping

Maximilian Gilles , Kai Furmans, and Rania Rayyes , *Member, IEEE*

**Abstract**—Learning-based robotic grasping systems typically rely on large-scale datasets for training. However, collecting such datasets in the real-world is both costly and time-consuming. Synthetic data generation data is a cost-effective alternative, but models trained solely on synthetic data often struggle with zero-shot real-world performance due to the large domain gap between synthetic and real-world data. To address this challenge of dataset costs against model performance, we propose an active learning framework designed for fast and sample-efficient sim-to-real domain adaptation. Our proposed learning framework uses synthetic data as initial knowledge base and incrementally adapts to the target data domain by selecting the most informative real-world data samples for further model training. For this purpose, we propose a novel, hybrid query strategy, MetaMVUC, which leverages multi-view uncertainty and metadata diversity. MetaMVUC assesses model uncertainty by comparing model predictions across multiple viewpoints, identifying samples with the highest uncertainty. Additionally, since robots in industry or logistics often operate in environments rich in metadata, MetaMVUC leverages this information to ensure diverse and well-distributed sample selection. Experimental results demonstrate the effectiveness of our proposed learning framework. With a limited annotation budget of 16 samples, a robot trained using MetaMVUC achieves a successful grasp rate of 87.7%. Increasing the budget to 40 samples improves grasp performance to 96.7%, outperforming the zero-shot sim-to-real by 17.4% and 26.4%, respectively.

**Index Terms**—Deep learning in grasping and manipulation, computer vision for automation, perception for grasping and manipulation.

## I. INTRODUCTION

**D**ETECTING and grasping objects in unstructured object heaps still represents a highly challenging task for robotic systems. Over recent years, data-driven deep learning methods have been proven to solve such tasks effectively and are being increasingly integrated in real-world robotic systems across various industries. In order to generalize well to unknown target data domains, such systems typically rely on training on large-scale datasets, which are both costly and time-consuming to

Received 27 September 2024; accepted 5 February 2025. Date of publication 19 February 2025; date of current version 6 March 2025. This article was recommended for publication by Associate Editor Y. Xiang and Editor M. Vincze upon evaluation of the reviewers' comments. The work of Rania Rayyes's was supported by the Baden-Württemberg Ministry of Science, Research and the Arts within Innovation Campus Future Mobility. (*Corresponding author: Maximilian Gilles.*)

The authors are with the Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany (e-mail: maximilian.gilles@kit.edu; kai.furmans@kit.edu; rania.rayyes@kit.edu).

This article has supplementary downloadable material available at <https://doi.org/10.1109/LRA.2025.3544083>, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3544083

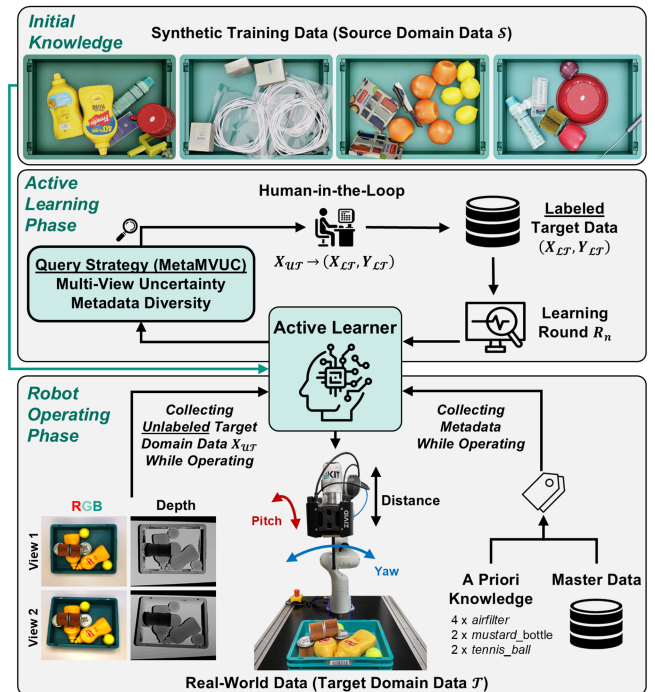


Fig. 1. Proposed AL framework for sample-efficient sim-to-real domain adaptation: Starting from synthetic data, the proposed MetaMVUC query strategy iteratively selects informative and diverse samples for incremental model adaptation. For a robot video, click <https://youtu.be/wsy9IRw19kQ>.

collect, representing a significant bottleneck in the deployment of learning-based robotic systems. One common approach to mitigate this issue of dataset costs is the use of synthetically generated datasets. However, models trained on simulation data often suffer from limited zero-shot performance due to a significant domain gap between the synthetic training and the real-world target domain. To address this challenge of dataset costs against model performance, we propose an active learning (AL) framework that enables sample-efficient and cost-effective sim-to-real domain adaptation for robotic vacuum grasping (cf. Fig. 1). We leverage the low cost of synthetic data for initial model training, creating a robust foundation of knowledge. Building on this cost-effective knowledge from simulation, our method iteratively queries a pool of unlabeled real-world data to identify the most informative samples, which are then labeled by an expert (human-in-the-loop) and used to incrementally adapt the learning model. Our approach assumes that some data samples provide more value to the training than others, given the current model state. By focusing on these informative samples,

our AL-based system can quickly adapt to new data domains, promoting both time- and sample-efficiency.

A key element of every AL system is its query strategy. Our proposed query function, MetaMVUC, uses multi-view uncertainty and metadata knowledge to select both informative and well-distributed data points for learning. Our approach leverages that inconsistencies in model predictions across different camera viewpoints serve as a reliable measure of model uncertainty – a concept demonstrated in [1] for the task of semantic segmentation of indoor scenes. While model uncertainty is commonly used to query informative samples, solely relying on it can lead to poor data distribution coverage due to a lack of diversity consideration. To address this, we recognize that metadata, such as object classes and quantities, is often available in real-world warehouse or production settings. Accordingly, we incorporate this prior knowledge to enhance sample diversity in our AL framework. In summary, our main contributions are as follows:

- We propose a novel AL framework for robotic grasping, enabling sample-efficient sim-to-real domain adaptation for instance segmentation and learning-based vacuum grasp detection.
- We propose a novel query strategy which leverages multi-view uncertainty and metadata diversity scoring.

To validate our proposed AL framework, we perform experiments on the MGNv2-Real dataset and in our physical robot cell, comparing our novel query strategy against SOTA methods and zero-shot sim-to-real transfer for instance segmentation and robotic grasping in clutter.

## II. RELATED WORK

### A. Robot Grasping in Clutter and Occlusion

In automation and logistics, data-driven vacuum grasping has become a standard due to its effectiveness in handling flat or tightly packed items, and narrow bins. Vacuum grasp detection is typically framed either as a pixel-wise regression task, producing grasp quality heatmaps for the entire image [2], [3], [4], or as a two-stage sample-based method [5], [6], where samples are ranked by their scores [7]. Cluttered and overlapping items present significant challenges, often leading to accidental multiple grasps or failures due to excessive contact forces [8], [9], [10]. Recent work addresses this challenge by various means, including the detection of a full object relationship tree [11], [12], [13], amodal instance segmentation masks [14], or by assessing the occlusion properties of individual objects without considering their adjacent relationships [4], [15].

*Proposed approach:* Our grasping method is based on [4]. Object instances are detected along with their occlusion property and semantic class. Vacuum grasp detection is framed as regression task of a grasp quality heatmap.

### B. Active Learning

The primary objective of AL is to optimize model performance while minimizing the costs associated with annotating training data. Over time, a variety of deep AL methods have been developed which can be categorized based on the availability of the unlabeled data (pool-based vs. stream-based) [16], the presence of initial model knowledge (cold start vs. warm start) [17],

or the type of query function employed [18]. Query functions are typically classified into diversity-based, uncertainty-based, and hybrid query strategies. Diversity-based query strategies aim to select a subset of samples that best represent the entire data distribution. The underlying assumption is that a good coverage of the data in the selected subset effectively filters out redundant samples or irrelevant outliers, leading to better sample efficiency. Prominent methods for diversity-based AL work by either jointly selecting a subset of samples that best covers the data in feature space [19], [20], [21], or by iteratively selecting the most representative samples [22], [23], [24]. Uncertainty-based query strategies focus on selecting samples where the model's predictions are most uncertain. These methods assume that data samples with high uncertainty contribute more to the training than those, where the model is already confident. Defining metrics for uncertainty in deep neural networks has been ongoing research for many years [25]. In the context of AL, uncertainty-based query strategies commonly employ methods based on information entropy [26], Bayesian models [27], ensembles [28], learning loss [29], and consistency [1], [30], [31], [32]. Hybrid query strategies combine the principles of uncertainty- and diversity-based sampling into a single method, aiming for both informative and at the same time representative samples [33], [34].

*Proposed Approach:* Our work explores the potential of using multiple camera viewpoints to query informative data samples for learning, proposing a novel method for multi-view uncertainty reasoning for instance segmentation and vacuum grasping tasks. In addition, we investigate the role of metadata in enhancing data diversity. By integrating both approaches — multi-view uncertainty and metadata diversity — into our novel hybrid query strategy, we aim to achieve sample-efficient learning for robotic grasping.

## III. METHOD

In our previous work MetaGraspNetv2 [4], we leveraged synthetic data generation to train both an object detection network  $f_{OD}$ , based on Mask-RCNN [35], and a vacuum grasp detection network  $f_{SC}$ , based on DeepLabv3 [36]. However, while synthetic data provides a cost-effective and time-efficient alternative to expensive real-world dataset collection, it often suffers from a significant sim-to-real gap and limited zero-shot performance. In this work, we address this critical challenge of balancing dataset costs against model performance via active domain adaptation. We design an AL framework (cf. Fig. 1) with a novel query strategy, **MetaMVUC**, that adapts  $f_{OD}$  and  $f_{SC}$ , pretrained on *source domain* data (synthetic), to the new *target domain* data (real-world) using only a small set of annotated target data.

### A. Problem Statement: Active Domain Adaptation

Let us denote  $\mathcal{S}$  for the source data domain, which contains sensor data  $X_{\mathcal{LS}}$  and their corresponding labels  $Y_{\mathcal{LS}}$ .  $\mathcal{T}$  represents the target data domain, which at the beginning only contains unlabeled data samples  $X_{\mathcal{UT}}$ . Throughout the AL process, selected samples  $X_{\mathcal{UT}}$  are annotated by a user  $X_{\mathcal{UT}} \rightarrow (X_{\mathcal{LT}}, Y_{\mathcal{LT}})$  and added to the labeled target domain data set  $\{(X_{\mathcal{LT}}, Y_{\mathcal{LT}})\}$  (cf. Fig. 1). In an active domain adaptation

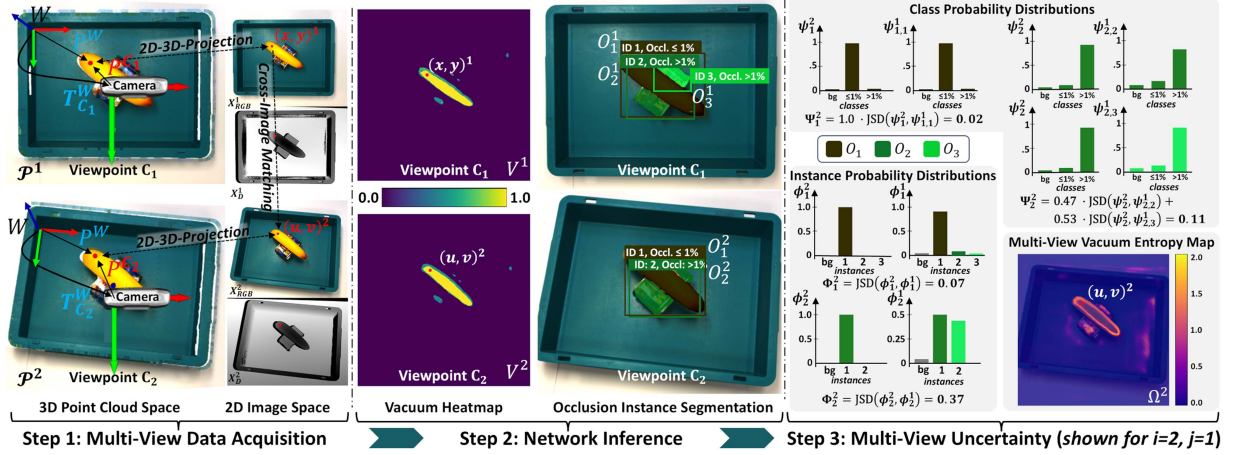


Fig. 2. Overview of proposed multi-view consistency-based uncertainty scoring pipeline. While the occlusion class is consistently predicted for both instances  $O_1^2$  and  $O_2^2$  across both reference viewpoint  $C_2$  and additional viewpoint  $C_1$ , the network struggles detecting the occluded object in  $C_1$ , resulting in a high value for instance consistency  $\Phi_2^2$ .

setup, the learning algorithm initially has access to  $\mathcal{S}$ , containing samples  $\{(X_{\mathcal{L}\mathcal{S}}, Y_{\mathcal{L}\mathcal{S}})\}$ . Following the pool-based AL approach, the static pool of unlabeled data samples  $X_{\mathcal{U}\mathcal{T}}$  is available from the start. In total, the learner has a query budget  $B \ll |X_{\mathcal{U}\mathcal{T}}|$ . For each learning round  $R_n$ ,  $n = 1 \dots N$ , and with a per-round query budget of  $\lfloor B/N \rfloor$ , the learner queries a subset of samples from the pool of  $X_{\mathcal{U}\mathcal{T}}$  and requests the user to annotate them  $X_{\mathcal{U}\mathcal{T}} \rightarrow (X_{\mathcal{L}\mathcal{T}}, Y_{\mathcal{L}\mathcal{T}})$ . Consequently, in each learning round  $R_n$ , the learner gains access to an incremented set of annotated target samples  $X_{\mathcal{L}\mathcal{T}}$ , while the total pool size of target samples  $X_{\mathcal{T}} = X_{\mathcal{L}\mathcal{T}} \cup X_{\mathcal{U}\mathcal{T}}$  remains constant. The goal of the learning algorithm is to improve object and grasp point detection in the new target data domain while keeping the number of annotated samples  $|X_{\mathcal{L}\mathcal{T}}|$  small, i.e., reducing labeling costs and speeding up learning.

### B. Multi-View Uncertainty Scoring

To estimate the uncertainty in the network prediction of  $f_{\text{SC}}$  and  $f_{\text{OD}}$ , we leverage the active vision capabilities of a robotic manipulator equipped with an arm-mounted camera. Our pipeline for multi-view consistency-based uncertainty scoring (MVUC) (cf. Fig. 2) is divided into the following steps: With its arm-mounted camera sensor the robot captures images from two different camera viewpoints. Network inference is performed on the data from each viewpoint. Reasoning about network uncertainty is done by matching predictions across both viewpoints.

**Step 1 – Multi-View Data Acquisition:** By moving the robot arm and its mounted camera, each grasp scene is observed from two different camera viewpoints  $C_k$  and  $k = 1, 2$ . Viewpoints are relative to a chosen world coordinate system  $W$  and defined by a transformation matrix  $T_{C_k}^W$ . The first viewpoint  $C_1$  is located overhead  $T_{C_1}^W$ , while the second viewpoint  $C_2$  has its pose  $T_{C_2}^W$  randomly chosen from a set of pre-configured poses, varying in distance, pitch, and yaw angles. At both camera viewpoints  $k = 1, 2$ , a color image  $X_{\text{RGB}}^k$  and a depth image  $X_D^k$  are captured, stored along with the camera pose  $T_{C_k}^W$  (cf. Step 1 in Fig. 2).

**Step 2 – Object and Grasp Detection Inference:** For both viewpoints  $C_k$  and  $k = 1, 2$ , object instances together with

their object class are detected given the color image  $X_{\text{RGB}}^k$ . Similar to our previous work [4],  $f_{\text{OD}}$  detects  $l$  object instances  $O_l^k = f_{\text{OD}}(X_{\text{RGB}}^k)$  pixel-wise along with their occlusion class: unoccluded, defined as less or equal to 1% occlusion, or occluded, defined as greater than 1% occlusion. Additionally, a vacuum grasp heatmap  $V^k$  is predicted for each viewpoint given the the depth image  $X_D^k$  (cf. Step 2 in Fig. 2). In line with [4], the heatmap encodes the vacuum sealability score for every pixel and is structured as a 3D tensor of shape  $[\mathcal{H}, \mathcal{W}, 25]$ , where  $\mathcal{H}$  and  $\mathcal{W}$  represent the height and width of the image, respectively, and 25 denotes the number of bins used to classify grasp scores. Each bin corresponds to a vacuum score within the range  $[0, 1]$  [4]. For every pixel, the model predicts a softmax distribution across these bins. The pixel-wise vacuum score is then obtained by taking the argmax of the softmax distribution:

$$V^k = \text{argmax}(\hat{V}^k) = \text{argmax}(\text{softmax}(f_{\text{SC}}(X_D^k))). \quad (1)$$

**Step 3 – Multi-View Uncertainty:** To reason about model uncertainty from multi-view predictions, pixels in images taken from both viewpoints corresponding to the same real-world point have to be matched. In our work, we achieve this by transforming both images into a common 3D space and identifying matching pixels through a nearest-neighbor search. Specifically, pixels  $(x, y)^j$  in image  $X^j$  and pixels  $(u, v)^i$  in image  $X^i$ , where index  $i$  represents the *reference* viewpoint  $C_i$  for  $i = 1, 2$  and index  $j$  the corresponding *additional* viewpoint  $C_j$  for  $j = 3 - i$ , can be projected to the same real-world point  $P^W$  (cf. Fig. 2 Step 1: 2D-3D Projection for  $i=2$  and  $j=1$ ). Using the depth image and the camera intrinsics, all pixels from image  $X^i$  and  $X^j$  are projected into 3D space. As a result, two point clouds  $\mathcal{P}$  are obtained, each containing points relative to the respective camera coordinate system  $C_i$  and  $C_j$  (cf. Fig. 2 Step 1). Given the camera poses  $T_{C_i}^W$  and  $T_{C_j}^W$ , all points from  $\mathcal{P}^i$  and  $\mathcal{P}^j$  are projected into the world coordinate system  $W$ . By aligning the points into the common coordinate system  $W$ , points in  $\mathcal{P}^j$  can be matched to points in  $\mathcal{P}^i$  based on their distance in 3D space. Knowing the correspondence between points in 3D space and pixels in image space, pixel-wise cross-image matching from image  $X^j$  to the



image  $X^i$  is achieved (cf. Fig. 2 Step 1: Cross-Image Matching):

$$(x, y)^j = f_{\text{cross}}((u, v)^i). \quad (2)$$

For the proposed multi-view uncertainty scoring, the consistency in network predictions across viewpoints is of interest. Given a reference viewpoint  $C_i$ , the mean softmax distribution of cross-matching vacuum predictions  $\hat{V}$  is computed. The pixel-wise entropy  $H$  of the mean softmax is then calculated to obtain the multi-view vacuum entropy map  $\Omega^i$ :

$$\Omega^i(u, v) = H \left( \frac{\hat{V}^i(u, v) + \hat{V}^j(x, y)}{2} \middle| (x, y)^j = f_{\text{cross}}((u, v)^i) \right) \quad (3)$$

(cf. Fig. 2 Step 3:  $\Omega^2$ ).  $\Omega^i$  is computed for both reference viewpoints  $i=1, 2$  and the *multi-view vacuum entropy* (MVVH) score is obtained as:

$$MVVH = \frac{1}{2} \left( \frac{1}{|\Omega^1|} \sum_{v=1}^{\mathcal{H}} \sum_{u=1}^{\mathcal{W}} \Omega^1 + \frac{1}{|\Omega^2|} \sum_{v=1}^{\mathcal{H}} \sum_{u=1}^{\mathcal{W}} \Omega^2 \right), \quad (4)$$

where  $\mathcal{W}$  and  $\mathcal{H}$  are the width and height of the image  $X_{\text{D}}^i$ .

Besides evaluating MVVH, a metric to assess the consistency of instance segmentation across viewpoints is proposed. Unlike semantic segmentation, instance segmentation groups pixels into coherent instances with a common identifier. The network is assumed to be confident if an object is consistently detected from both viewpoints. To quantify instance consistency  $\Phi_l^i = \text{JSD}(\phi_l^i, \phi_l^j)$  for a detected object instance  $O_l^i$  in a reference viewpoint  $C_i$ , the Jensen-Shannon divergence (JSD) is used. It measures the similarity between corresponding instance probability distributions  $\phi_l^i$  and  $\phi_l^j$ . To compute  $\phi_l^j$ , cross-projected pixels associated with instance predictions in  $C_j$  are counted, given the detection  $O_l^i$  in  $C_i$ . If there is no detection in  $C_j$ , these pixels are categorized as background. The counts are normalized by the sum of the cross-matching pixels and sorted in descending order, yielding the probability distribution  $\phi_l^j$ . See Fig. 2 for an example: while instance  $O_1^2$  is consistently predicted across both reference viewpoint  $C_2$  and additional viewpoint  $C_1$ , resulting in a low value for  $\Phi_1^2$ , object instance  $O_2^2$  is incorrectly detected as two instances  $O_2^1$  and  $O_3^1$  in  $C_1$ , leading to a high value for  $\Phi_2^2$ .  $\Phi_l^i$  is computed for all object instances  $O_l^i$  for both reference viewpoints  $i=1, 2$ , resulting in the *multi-view instance consistency* (MVIC) score as:

$$MVIC = \frac{1}{2} \left( \frac{1}{\mathcal{N}} \sum_{l=1}^{\mathcal{N}} \Phi_l^1 + \frac{1}{\mathcal{M}} \sum_{l=1}^{\mathcal{M}} \Phi_l^2 \right), \quad (5)$$

where  $\mathcal{N}$  and  $\mathcal{M}$  are the total number of detected object instances in image  $X_{\text{RGB}}^1$  and  $X_{\text{RGB}}^2$ , respectively.

We further extend to quantify the multi-view consistency of class probability distributions. Given a reference viewpoint  $C_i$ , the multi-view class consistency  $\Psi_l^i$  for an object instance  $O_l^i$  is calculated using JSD for measuring the similarity between corresponding class probability distributions  $\psi$ .  $\psi_l^i$  represents the class probability distribution for an object instance  $O_l^i$  in  $C_i$ , while  $\psi_{l,l'}^j$  characterizes the class probability distributions of corresponding instance predictions  $l'$  for  $O_l^i$  in  $C_j$ . To obtain  $\Psi_l^i$ , similarity measurements  $\text{JSD}(\psi_l^i, \psi_{l,l'}^j)$  are summed up and

weighted by pixel-area:

$$\Psi_l^i = \sum_{l'} w_{l,l'} \cdot \text{JSD}(\psi_l^i, \psi_{l,l'}^j). \quad (6)$$

See Fig. 2 for an occlusion class detection example: The network predicts that instance  $O_1^2$  in reference viewpoint  $C_2$  and its corresponding instance  $O_{1,1}^1$  in viewpoint  $C_1$  are both unoccluded, resulting in high consistency between  $\psi_1^2$  and  $\psi_{1,1}^1$  and a low value for  $\Psi_1^2$ . Similarly, for detection  $O_2^2$  in  $C_2$ , the network predicts that both corresponding predictions in  $C_1$ ,  $O_{2,2}^1$  and  $O_{2,3}^1$ , are occluded, leading to a low value for  $\Psi_2^2$ .  $\Psi_l^i$  is computed for all instances  $O_l^i$  in both reference viewpoints  $i=1, 2$ , resulting in the *multi-view class consistency* (MVCC) score:

$$MVCC = \frac{1}{2} \left( \frac{1}{\mathcal{N}} \sum_{l=1}^{\mathcal{N}} \Psi_l^1 + \frac{1}{\mathcal{M}} \sum_{l=1}^{\mathcal{M}} \Psi_l^2 \right), \quad (7)$$

where  $\mathcal{N}$  and  $\mathcal{M}$  are the total number of detected object instances in image  $X_{\text{RGB}}^1$  and  $X_{\text{RGB}}^2$ , respectively.

In our AL framework, the proposed scoring functions are applied to all samples in  $X_{\text{UT}}$ . Results are sorted in descending order, yielding individual rankings  $\nu_{\Omega}$ ,  $\nu_{\Phi}$ , and  $\nu_{\Psi}$  for MVVH, MVIC, and MVCC, respectively.

### C. Metadata Diversity Scoring

Relying solely on uncertainty-based query methods can result in redundant sample selection [33]. To address this, we incorporate metadata for diversity scoring. However, it is important to note that metadata alone does not provide ground truth for training  $f_{\text{OD}}$  and  $f_{\text{SC}}$ . Our algorithm queries diverse and balanced samples using prior knowledge (*metadata*) of each object's class in the grasp scene, the number of instances, and the primary material of an object - information typically maintained in warehouse or production settings. The algorithm iterates through all unlabeled samples  $X_{\text{UT},j}$ ,  $j=1 \dots |X_{\text{UT}}|$ , ranking each sample based on its contribution to creating a diverse and well-balanced query set in terms of class distribution  $\rho_C$ , number of instances per scene distribution  $\rho_n$ , and instance material distribution  $\rho_M$ . The diversity contribution of a sample  $X_{\text{UT},j}$  is quantified using JSD between the query-set distribution  $\rho$ , comprising of the labeled samples plus the considered sample  $\{X_{\text{LT}}, X_{\text{UT},j}\}$ , and a uniform distribution  $\tilde{\rho}$ . The uniform distribution is used to promote a well-balanced query-set. The *metadata diversity score* (MDDS) is defined as:

$$MDDS = \frac{1}{3} \sum_k^{\{C,n,M\}} \text{JSD}(\rho_k(X_{\text{LT}} + X_{\text{UT},j}), \tilde{\rho}_k). \quad (8)$$

In our AL framework, MDDS is applied to all samples in  $X_{\text{UT}}$ . The resulting scores are ranked in ascending order to obtain a metadata diversity ranking  $\nu_{\zeta}$ .

### D. Hybrid Query Strategy: MetaMVUC

To leverage both diversity- and uncertainty-based query strategies, the proposed multi-view uncertainty and metadata diversity scoring metrics are integrated into a hybrid query strategy named **MetaMVUC**. The Borda rule method [37] is employed to achieve an aggregated ranking across all metrics,

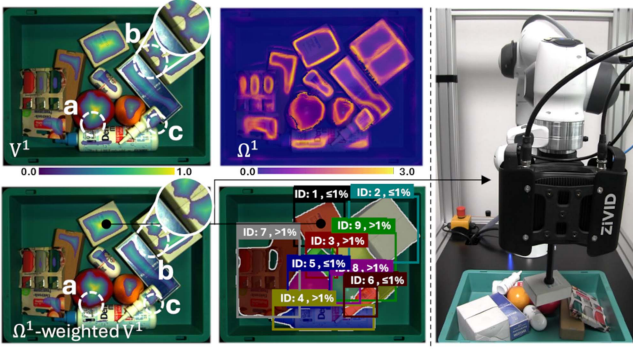


Fig. 3. Artifacts in the vacuum heatmap  $V^1$  can be effectively filtered out (cf. white circles a, b, and c in  $\Omega^1$ -weighted  $V^1$ ).

where each metric –  $MVH$ ,  $MVIC$ ,  $MVCC$ , and  $MDDS$  – acts as voter  $v \in \mathbb{V} = \{\Omega, \Phi, \Psi, \zeta\}$ . Following [37], the rank of the  $j$ -th sample  $X_{UT,j}$  under the voter  $v$  is denoted as  $\text{rk}(X_{UT,j}, \nu_v)$ . For example,  $\text{rk}(X_{UT,j^*}, \nu_\Omega) = 1$  expresses that the sample  $j^*$  is the highest-ranked sample in the ranking  $\nu_\Omega$  according to the  $MVH$  metric. The Borda rule assigns each sample in  $X_{UT}$  a score based on its aggregated ranking. Specifically, the Borda score for each sample  $X_{UT,j} \in X_{UT}$  is calculated as the  $\lambda_v$  weighted voter-specific score  $|X_{UT}| - \text{rk}(X_{UT,j}, \nu_v)$ , and summed over all voters  $v \in \mathbb{V}$ :

$$\text{Borda}(X_{UT,j}) = \sum_{v \in \mathbb{V}} \lambda_v \cdot (|X_{UT}| - \text{rk}(X_{UT,j}, \nu_v)), \quad (9)$$

with the weights chosen as:  $\lambda_\Phi = \lambda_\Omega = \lambda_\Psi = \lambda_\zeta = 1$ .

This scoring system identifies samples that achieve broad agreement across all metrics. Given the per-round query budget  $b_n$ , the top- $b_n$  highest-ranked samples based on their Borda scores are selected. Once annotated, these samples are added to the training dataset and used for iterative model adaptation to the real target domain  $\mathcal{T}$ .

#### E. $\Omega$ -Weighted Occlusion-Aware Vacuum Grasping

At inference time, either between learning rounds or after the AL process is complete, the proposed multi-view vacuum entropy map  $\Omega^1$  for reference viewpoint  $i=1$  (cf. Eq (3)) can improve grasp detection. Our  $\Omega$ -weighted vacuum grasping algorithm prioritizes areas of lower uncertainty, which are generally more reliable for grasping, over areas of higher uncertainty, which often correspond to edges or object boundaries (cf. Fig. 3  $\Omega^1$ ). This is achieved by convolving the vacuum grasp heatmap  $V^1$  with the normalized and inverted vacuum entropy map  $\Omega^1$  using convolution operation (cf. Fig. 3  $\Omega^1$ -weighted  $V^1$ ). This heatmap is integrated into our previous occlusion-aware grasp detection algorithm [4]. Specifically, unoccluded objects are prioritized for grasping over occluded ones unless an occluded grasp's score exceeds the highest unoccluded grasp score by a factor of two. This exception accounts for scene configurations where all unoccluded objects might be challenging to grasp, making a high-scoring occluded grasp more practical.

## IV. EXPERIMENTS

In all our experiments, the proposed AL framework was used. We evaluate the effectiveness of our proposed query-strategy, MetaMVUC, for real-world tasks associated to robotic grasping. The primary research question is: How does MetaMVUC (MetaMVUC) compare to the diversity-based AL method core-set [19] (CORE), the hybrid AL method CLUE [33] (CLUE), and random sampling (RAND) in improving occlusion-aware object detection, semantic object detection, grasp performance and order picking performance?

### A. Experimental Setup

**Datasets and Real-World Setup:** Experiments are performed on the real data split of MetaGraspNetv2 dataset [4] (MGNv2-Real) and in our physical robot cell (cf. Fig. 3). The MetaGraspNetv2 dataset, which focuses on object grasping in cluttered scenes, offers practical advantages for our study. Initially collected using the same setup as our AL experiments, it enables us to replicate realistic pool-based AL conditions in our robot cell, without the need for extensive labeling. The data is already pre-annotated for object detection and vacuum grasping by expert users who are familiar with the system and the object set. We select 33 objects from the MGNv2-Real dataset, referred to as *seen* objects, and introduce 23 novel objects (*unseen* objects). MGNv2-Real was filtered to include only scenes containing the seen objects, resulting in 233 samples. They were split into two sets: 70% form the pool set (MGNv2-Pool), and the remaining 30% form the test set (MGNv2-Test).

**Training Details:** As a starting point, vacuum grasp detection network  $f_{SC}$  and instance segmentation network  $f_{OD}$  are pre-trained for 20 epochs on the synthetic dataset MGNv2-Sim [4]. The Adam optimizer with a learning rate of 0.001 is used for  $f_{OD}$ , and the SGD optimizer with a learning rate of  $1 \times 10^{-5}$  is used for  $f_{SC}$ . All network layers are trainable. A batch size of 4 is chosen and the per-round query budget is set to  $b_n = 4$  samples. Each sample is annotated for both viewpoints  $C_1$  and  $C_2$ . We train our models for  $N = 10$  learning rounds, utilizing different query strategies: MetaMVUC, CLUE, CORE, and RANDOM. Each learning round  $R_n$  consists of 20 training epochs (full iteration through training set), and training is continued from the previous learning round. Each experiment is repeated five times using different random seeds. Reported results are the mean averages of these runs.

**Metrics:** For occlusion instance segmentation and semantic instance segmentation, we evaluate performance using the mean Average Precision (mAP) metric across all classes at an instance mask intersection over union (IoU) threshold of 0.75. Additionally, we perform statistical testing using two-tailed paired  $t$ -tests after each learning round and present the results for all possible pairwise comparisons among MetaMVUC, CLUE, CORE, and RANDOM in the form of a Pairwise Penalty Matrix (PPM) (cf. Fig. 4), as proposed by [34], [38]. The PPM reflects statistical differences between query strategies, aggregated across learning rounds. Following the protocol outlined in [38], we adopt a confidence level of 90% and define a penalty score of  $\frac{1}{N} = 0.1$ , where  $N$  is the number of learning rounds. For each learning round  $R_n$ , where  $n = 1 \dots N$ , if the  $t$ -value of method  $i$  is above the confidence interval (method  $i$  outperforms method  $j$  at learning

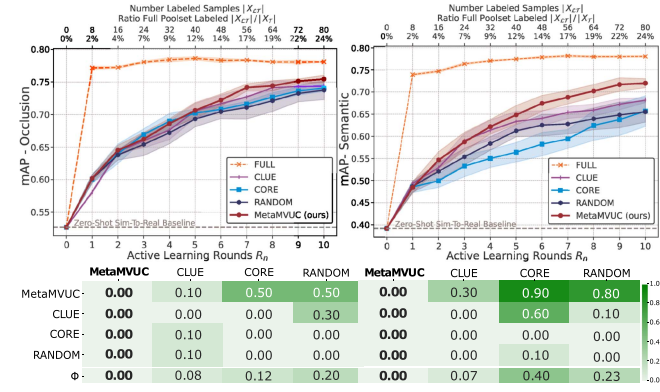


Fig. 4. Results on MGNv2-Test with standard deviation indicated in shaded area (*top row*). The corresponding Pairwise Penalty Matrices (PMM) are shown in the *bottom row*.

round  $R_n$ ), a penalty is added to the cell at position  $(i, j)$  in the PMM, with the first index representing the row and the second the column. Conversely, if the  $t$ -value is below the confidence interval (method  $j$  outperforms method  $i$  at learning round  $R_n$ ), a penalty is added to the cell at position  $(j, i)$ . Consequently, the method with the lowest column-wise average is considered the best performer, while the method with the highest average is considered the worst. In our physical robot experiments, grasp performance is measured by the number of successful grasps over the total number of grasps attempts  $R_{\text{grasp}}$  (successful grasp rate) and the number of successfully cleared objects over the total number of objects  $R_{\text{obj}}$  (cleared object rate). Order picking performance is measured by the average order fulfillment rate  $F_{\text{order}}$  and the classification precision,  $P_{\text{class}}$ , which represents the number of correctly classified objects among all successfully grasped objects.

### B. Object Instance Segmentation Results on MGNv2-Test

Instance segmentation performance is evaluated on the MGNv2-Test dataset, using MGNv2-Pool as the query pool  $X_{\mathcal{T}}$ . Results are reported in Fig. 4 for the task of detecting instances together with their occlusion class and their semantic class for different query set sizes  $|X_{\mathcal{T}}|$  and learning rounds  $R_n$ . The FULL line in Fig. 4 plots results over learning rounds when the entire query pool  $X_{\mathcal{T}}$  is used for training, serving as an upper baseline for comparison.

For the task of occlusion instance segmentation, MetaMVUC outperforms CORE, CLUE, and RAND in the later AL rounds ( $n \geq 5$ ). For semantic instance segmentation, the reported results in Fig. 4 show that MetaMVUC outperforms all other methods by a substantial margin, with mAP performance improvements of 6.3% and 6.5% over RANDOM, 7.8% and 6.3% over CORE, and 4.3% and 3.8% over CLUE for learning round  $R_8$  and  $R_{10}$ , respectively. The pairwise penalty matrix (PPM) in Fig. 4 *bottom row* highlights MetaMVUC's clear superiority over CORE and RANDOM sampling, with an advantage over CLUE for both occlusion and semantic instance segmentation. For ablation studies, the proposed multi-view uncertainty scoring mechanism is compared against Bayesian methods and the impact of diversity sampling within MetaMVUC is evaluated. To this

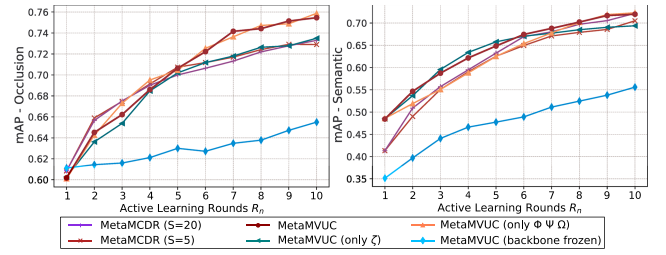


Fig. 5. Ablation studies for MetaMVUC.

end, we replace the multi-view uncertainty of MetaMVUC with Monte Carlo dropout sampling (MCDR), a well-established approximation for Bayesian inference [39]. In the following, we will refer to this as MetaMCDR. Results for semantic instance segmentation in Fig. 5 show that the combination of both scoring strategies, multi-view uncertainty ( $\Omega, \Phi, \Psi$ ) and metadata diversity ( $\zeta$ ), within MetaMVUC leads to consistently strong performance across all learning rounds, while for occlusion detection additional diversity scoring provides neither a significant benefit nor drawback (cf. *orange* and *dark red* line). Our results for MetaMCDR indicate that multi-view sampling outperforms MCDR as an uncertainty signal for occlusion detection. For semantic instance segmentation, MetaMCDR matches the performance of MetaMVUC in later active learning rounds ( $n \geq 6$ ) and shows lower performance in the earlier rounds ( $n < 6$ ). Importantly, MCDR requires  $S=20$  network inferences per image for achieving competitive performance with MetaMVUC, resulting in high computational costs and long query times. Notably, when the number of samples is reduced to  $S=5$ , MCDR exhibits significantly lower performance. Studies for MetaMVUC on freezing the backbone layers during training (cf. Fig. 5 backbone frozen) reveal a significant performance drop for both tasks compared to using fully trainable network layers.

*Summary:* These findings demonstrate that MetaMVUC adapts quickly and sample-efficiently to new target domain data outperforming a diversity-based AL method (CORE), a hybrid AL method (CLUE), and a random baseline (RANDOM), as shown by the PMMs in Fig. 4. While (FULL) yields the best performance, as expected, MetaMVUC achieves strong results with only a fraction of the data (see Fig. 4). Ablation studies in Fig. 5 show that MetaMVUC can also be a valid approach when metadata or multi-view sensing is not available.

### C. Grasping Performance in Robot Cell

In our robot experiments, we evaluate the real-world performance of the  $\Omega$ -weighted occlusion-aware vacuum grasping method (cf. Section III-E), trained with different query strategies: MetaMVUC, CLUE, CORE, and RANDOM. We compare MetaMVUC with CLUE, CORE, and RANDOM to assess its impact on sample efficiency  $|X_{\mathcal{T}}|$  and system performance in terms of grasp performance ( $R_{\text{grasp}}$  and  $R_{\text{obj}}$ ) and object detection ( $F_{\text{order}}$  and  $P_{\text{class}}$ ). For all query strategies, we conduct physical robot grasping experiments after learning rounds  $R_2$ ,  $R_5$ , and  $R_{10}$ . For comparative analysis, we also perform experiments *without* AL ( $R_0$ ), representing a zero-shot sim-to-real baseline (Zero-Shot). Experiments are performed on two object sets:



TABLE I  
REAL ROBOT PERFORMANCE [%] WITH STD IN PARENTHESES ACROSS  
LEARNING ROUNDS  $R_n$  AND OBJECT SETS

Method	$R_n$	Seen Objects				Unseen Objects	
	$n$	$R_{\text{grasp}}$	$R_{\text{obj}}$	$P_{\text{class}}$	$F_{\text{order}}$	$R_{\text{grasp}}$	$R_{\text{obj}}$
Zero-Shot		70.3 (18.3)	82.0 (11.7)	52.6 (15.5)	42.0 (9.2)	79.7 (9.5)	91.0 (5.4)
MetaMVUC	2	87.7 (10.3)	94.0 (4.9)	68.2 (13.0)	63.0 (11.6)	81.6 (11.5)	93.0 (7.8)
	5	96.7 (6.7)	100.0 (0.0)	71.0 (12.0)	70.0 (12.5)	89.7 (9.3)	97.0 (4.6)
	10	96.4 (6.0)	99.0 (3.0)	81.0 (12.0)	77.0 (12.5)	86.5 (12.0)	95.0 (6.7)
CLUE	2	76.1 (7.7)	87.0 (4.6)	70.7 (21.5)	59.0 (17.9)	68.2 (16.8)	83.0 (11.0)
	5	85.9 (13.8)	93.0 (7.8)	69.6 (14.2)	64.0 (15.8)	88.7 (10.4)	95.0 (5.0)
	10	90.2 (8.4)	96.0 (4.9)	75.0 (11.0)	71.0 (9.9)	86.1 (9.3)	95.0 (5.0)
CORE	2	84.6 (14.0)	94.0 (8.0)	51.9 (22.1)	47.0 (21.1)	79.2 (15.1)	90.0 (8.9)
	5	87.7 (8.8)	96.7 (4.7)	60.9 (13.3)	57.0 (14.9)	81.5 (12.4)	93.0 (6.4)
	10	89.0 (13.5)	95.0 (9.2)	68.4 (12.7)	64.0 (14.3)	86.9 (10.8)	94.0 (6.6)
RANDOM	2	89.0 (11.6)	95.0 (5.0)	58.1 (12.6)	52.0 (14.8)	82.4 (11.8)	93.0 (7.8)
	5	91.9 (11.7)	97.0 (4.6)	67.3 (13.4)	63.0 (12.5)	89.9 (11.5)	97.0 (6.4)
	10	98.3 (5.0)	100.0 (0.0)	70.0 (11.5)	68.0 (11.4)	86.1 (8.3)	96.0 (4.9)
Color Codes: Rank #1, Rank #2, Rank #3, Rank #4							
Please note: Methods are ranked separately for each learning round $R_2$ , $R_5$ , and $R_{10}$ .							

seen objects included in the pool set  $X_T$ , and unseen objects, not part of the pool set and novel at test time. Each run starts with a cluttered scene of 10 objects, and the robot is tasked to grasp all objects the scene. For the seen object split, an additional order picking task is introduced: In addition to emptying the whole scene, the robot must also predict the semantic class of each grasped object and place it in one of five designated bins, mimicking a typical order-picking task. For unseen objects, this classification task is omitted, and the robot is instructed to place the objects randomly in the bins. After two failed grasp attempts per object and run, the object is manually removed. Reported numbers in Table I are averaged over 10 runs with different random seeds.

For seen objects, which represent the typical scenario for pool-based AL use-cases, the results in Table I show that MetaMVUC consistently excels in both the standard task of grasping all objects in the scene and the more complex order-picking task, as indicated by the frequent top rankings (shown in darker green) in Table I. Notably, for the order-picking task, which requires both reliable grasping and object detection, MetaMVUC outperforms all other methods in terms of order fulfillment  $F_{\text{order}}$  and correctly classified picked items  $P_{\text{class}}$ . When focusing solely on seen object grasping, MetaMVUC achieves overall high success rates in both successful grasp rate  $R_{\text{grasp}}$  and object clearance rate  $R_{\text{obj}}$ , demonstrating competitive performance compared to RANDOM after learning rounds  $R_2$  and  $R_{10}$ , and outperforming it for learning round  $R_5$ . However, when comparing MetaMVUC to CLUE and CORE, MetaMVUC clearly outperforms both by a substantial margin across all metrics. The results show that with just 16 ( $R_2$ ) annotated data samples out of 324, our robot achieves successful grasp rates of 87.7%, object clearance rates of 94.0% and order fulfillment rates of 63.0%, improving performance by 17.4%, 12.0% and 21.0% compared to the zero-shot sim-to-real transfer. Increasing the budget to 40 samples ( $R_5$ ) improves the average successful grasp rate to 96.7%. After 10 learning rounds, which is equivalent to 80 annotated images ( $R_{10}$ ), our robot trained using MetaMVUC correctly classifies 81.0% of picked items, improving the order fulfillment rate by 35.0% compared to Zero-Shot and by 9.0% to RANDOM. A qualitative analysis in Fig. 6 supports these findings: as the models are exposed to more training data, performance improves and uncertainty decreases, as shown by increasingly well-defined  $\Omega$ -maps.

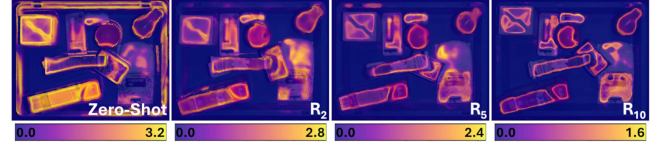


Fig. 6. Overview of  $\Omega$ -Heatmaps across learning rounds.

For unseen objects, as expected, a general decline in performance compared to seen objects is observed. Within this trend, MetaMVUC and RANDOM achieve competitive performance, outperforming both CORE and CLUE. Notably, the best grasp performance in terms of  $R_{\text{grasp}}$  and  $R_{\text{obj}}$  is achieved after 5 learning rounds, with a decline observed after learning round  $R_{10}$ . This observation highlights a general challenge with pool-based AL approaches, where the static data pool can limit the model's ability to generalize.

**Error Analysis:** We identified two primary error types causing failed grasp attempts: vacuum seal failure (Type I) and excessive contact wrench forces (Type II). These errors occur at different stages of the grasping process: Type I occurs when the suction cup does not establish an initial vacuum seal, while Type II occurs during lifting, after an initial successful contact. For MetaMVUC, 73% of observed failures with seen objects can be attributed to Type II errors. They often occur when dealing with elongated objects due to shortcomings in the center of mass heuristic integrated into the grasp detection pipeline proposed in [4], leading to grasp points located too far from the object's center of mass. The remaining 27% of failures are Type I errors, mainly caused by poor vacuum seal predictions or unexpected object movement during the initial contact phase. For unseen objects, Type I errors account for 60% of all failures, primarily due to poor predictions from the grasp network. The remaining 40% are Type II errors, caused by failures in the center of mass heuristic or inaccuracies in occlusion detection, resulting in excessive contact wrench forces from overlying objects.

**Summary:** Table I shows that zero-shot sim-to-real transfer performs well, despite the significant gap between simulation and the real world. However, an average successful grasp rate of 70.3% remains insufficient for real-world applications and highlights the importance of this work in effectively bridging that gap. As previous studies have also noted [23], [40], RANDOM, despite its simplicity, serves as a strong baseline for deep AL methods. This can be attributed to the fact that many datasets, including the MGNv2-Real dataset [4] used as our pool set, are well-balanced from the outset. Such balance reduces the advantage of sophisticated active learning methods over random selection, as the pool set is inherently informative. However, in real-world applications, a typical pool set is likely to contain more redundant data samples. In the challenging order-picking task, which requires both reliable grasping and object detection, MetaMVUC clearly outperforms RANDOM, as well as CORE and CLUE. This outcome is expected, given the strong performance of MetaMVUC in semantic instance segmentation (cf. Fig. 4 right). Notably, the commonly used diversity-based AL method CORE shows significantly lower performance compared to MetaMVUC. Although not the primary focus of AL, MetaMVUC also demonstrates strong performance on grasping unseen objects.

## V. CONCLUSION

In this letter, we design a novel AL framework with a novel hybrid query strategy, MetaMVUC, for robotic grasping learning that enables sample-efficient training in the real-world. Given the high costs associated with data collection and annotation, sample-efficient robot learning systems are crucial. Our AL framework leverages synthetic data as initial model knowledge. By combining multi-view uncertainty with metadata diversity scoring, the proposed MetaMVUC query strategy then selects the most informative and representative real-world samples for learning. Our experiments demonstrate that MetaMVUC is highly effective for learning occlusion and semantic instance segmentation tasks, as well as complex real-world grasping tasks, all while keeping the number of annotated samples small. Especially for the challenging task of semantic instance segmentation, MetaMVUC shows substantial improvements across all learning rounds. While performance improves with more data, the level of improvement is controlled by the annotation budget. Remarkably, with only 16 samples, a robot trained using MetaMVUC achieves a successful grasp rate of 87.7%. Increasing the budget to 40 samples (12% of the pool set) improves grasp performance to 96.7%, outperforming the zero-shot sim-to-real by 17.4% and 26.4%, respectively.

Based on our results, we believe that our work can contribute to the development of low-cost real-world robots.

## REFERENCES

- [1] Y. Siddiqui, J. Valentin, and M. Niessner, "ViewAL: Active learning with viewpoint entropy for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9430–9440.
- [2] H. Cao, H.-S. Fang, W. Liu, and C. Lu, "SuctionNet-1billion: A large-scale benchmark for suction grasping," *IEEE Robot. Automat. Lett.*, vol. 6, no. 4, pp. 8718–8725, Oct. 2021.
- [3] P. Jiang et al., "Learning suction graspability considering grasp quality and robot reachability for bin-picking," *Front. Neurobot.*, vol. 16, 2022, Art. no. 806898.
- [4] M. Gilles et al., "MetaGraspNetV2: All-in-one dataset enabling fast and reliable robotic bin picking via object relationship reasoning and dexterous grasping," *IEEE Trans. Automat. Sci. Eng.*, vol. 21, no. 3, pp. 2302–2320, Jul. 2024.
- [5] 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. Automat.*, 2018, pp. 5620–5627.
- [6] H. Zhang, J. Peeters, E. Demeester, and K. Kellens, "Deep learning reactive robotic grasping with a versatile vacuum gripper," *IEEE Trans. Robot.*, vol. 39, no. 2, pp. 1244–1259, Apr. 2023.
- [7] K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber, "A survey on learning-based robotic grasping," *Curr. Robot. Reports*, vol. 1, pp. 239–249, 2020.
- [8] Z. Pan, A. Zeng, Y. Li, J. Yu, and K. Hauser, "Algorithms and systems for manipulating multiple objects," *IEEE Trans. Robot.*, vol. 39, no. 1, pp. 2–20, Feb. 2023.
- [9] L. Li, A. Cherouat, H. Snoussi, and T. Wang, "Grasping with occlusion-aware ally method in complex scenes," *IEEE Trans. Automat. Sci. Eng.*, early access, Aug. 01, 2024, doi: [10.1109/TASE.2024.3434610](https://doi.org/10.1109/TASE.2024.3434610).
- [10] Y. Xia et al., "TARGO: Benchmarking target-driven object grasping under occlusions," 2024, *arXiv:2407.06168*.
- [11] 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*, 2018, pp. 118–125.
- [12] G. Zuo, J. Tong, H. Liu, W. Chen, and J. Li, "Graph-based visual manipulation relationship reasoning network for robotic grasping," *Front. Neurobot.*, vol. 15, 2021, Art. no. 719731.
- [13] 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.*, 2022, pp. 1404–1410.
- [14] S. Back et al., "Unseen object amodal instance segmentation via hierarchical occlusion modeling," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 5085–5092.
- [15] W. Liu, J. Hu, and W. Wang, "A novel camera fusion method based on switching scheme and occlusion-aware object detection for real-time robotic grasping," *J. Intell. Robot. Syst.*, vol. 100, no. 3, pp. 791–808, 2020.
- [16] D. Cacciarelli and M. Kulahci, "Active learning for data streams: A survey," *Mach. Learn.*, vol. 113, no. 1, pp. 185–239, 2024.
- [17] Q. Jin et al., "Cold-start active learning for image classification," *Inf. Sci.*, vol. 616, pp. 16–36, 2022.
- [18] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: A state of the art," *Artif. Intell. Rev.*, vol. 56, no. 4, pp. 3005–3054, 2023.
- [19] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *Proc. 6th Int. Conf. Learn. Representations*, 2018, pp. 1–13. [Online]. Available: <https://openreview.net/pdf?id=H1aIuk-RW>
- [20] S. Agarwal, H. Arora, S. Anand, and C. Arora, "Contextual diversity for active learning," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 137–153.
- [21] C. Guo, B. Zhao, and Y. Bai, "DeepCore: A comprehensive library for coresets selection in deep learning," in *Proc. Int. Conf. Database Expert Syst. Appl.*, 2022, pp. 181–195.
- [22] D. Gissin and S. Shalev-Shwartz, "Discriminative active learning," 2019, *arXiv:1907.06347*.
- [23] B. Wei et al., "Discriminative active learning for robotic grasping in cluttered scene," *IEEE Robot. Automat. Lett.*, vol. 8, no. 3, pp. 1858–1865, Mar. 2023.
- [24] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5971–5980.
- [25] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 1513–1589, 2023.
- [26] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, CS Tech. Rep. TR1648, 2009.
- [27] A. Kirsch, J. van Amersfoort, and Y. Gal, "BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 7026–7037.
- [28] W. H. Beluch, T. Genewein, A. Nurnberger, and J. M. Kohler, "The power of ensembles for active learning in image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9368–9377.
- [29] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 93–102.
- [30] I. Elezi, Z. Yu, A. Anandkumar, L. Leal-Taixé, and J. M. Alvarez, "Not all labels are equal: Rationalizing the labeling costs for training object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14472–14481.
- [31] S. A. Golestaneh and K. M. Kitani, "Importance of self-consistency in active learning for semantic segmentation," in *Proc. 31th Brit. Mach. Vis. Conf.*, 2020, pp. 1–16. [Online]. Available: <https://www.bmvc2020-conference.com/assets/papers/0010.pdf>
- [32] W. Yu, S. Zhu, T. Yang, and C. Chen, "Consistency-based active learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 3950–3959.
- [33] V. Prabhu, A. Chandrasekaran, K. Saenko, and J. Hoffman, "Active domain adaptation via clustering uncertainty-weighted embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8485–8494.
- [34] J. T. Ash et al., "Deep batch active learning by diverse, uncertain gradient lower bounds," in *Proc. 8th Int. Conf. Learn. Representations*, 2020, pp. 1–26. [Online]. Available: <https://openreview.net/pdf?id=ryghZJBKPS>
- [35] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
- [36] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
- [37] D. Burka, C. Puppe, L. Szepesváry, and A. Tasnádi, "Voting: A machine learning approach," *Eur. J. Oper. Res.*, vol. 299, no. 3, pp. 1003–1017, 2022.
- [38] Y. Ji, D. Kaestner, O. Wirth, and C. Wressnegger, "Randomness is the root of all evil: More reliable evaluation of deep active learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 3932–3941.
- [39] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1050–1059.
- [40] S. Mittal, J. Niemeijer, J. P. Schäfer, and T. Brox, "Best practices in active learning for semantic segmentation," in *Proc. DAGM German Conf. Pattern Recognit.*, 2024, pp. 427–442.