MetaMVUC: Sim-to-Real Active Domain Adaptation based on Multi-View Uncertainty and Metadata for Sample-Efficient Robotic Grasping

Maximilian Gilles
Karlsruhe Institute of Technology
76131 Karlsruhe, Germany
Email: maximilian.gilles@kit.edu

Kai Furmans
Karlsruhe Institute of Technology
76131 Karlsruhe, Germany
Email: kai.furmans@kit.edu

Rania Rayyes
Karlsruhe Institute of Technology
76131 Karlsruhe, Germany
Email: rania.rayyes@kit.edu

***Abstract*—Good generalization of learning-based robotic grasping systems to unknown target data domains requires training on large-scale datasets. However, collecting such datasets is very costly and time-consuming. In addition, these systems often have limited zero-shot performance, especially when they are trained on synthetic data. To overcome these limitations of passive robot learning, we establish a novel active learning framework to enable fast and sample-efficient adaptation to a new real-world target data domain. Our proposed learning framework uses synthetic data as a starting point and then selects the most informative real-world target data samples for incremental domain adaptation. For this purpose, we propose a novel query strategy, MetaMVUC, which leverages multi-view uncertainty and metadata diversity. Our strategy uses multiple viewpoints of the scene to reason about model uncertainty by matching predictions across viewpoints and identifying samples with the highest uncertainty. Additionally, since robots in industry or logistics often operate in environments rich in metadata, MetaMVUC utlizes this metadata to sample diverse and well-distributed samples. Experimental results on the MGNv2 dataset and in our physical robot cell clearly demonstrate the effectiveness and the robustness of our proposed learning framework built upon MetaMVUC. Real grasp experiments show that with only 16 out of 324 annotated data samples, our system achieves successful grasp rates of more than 87% for seen objects and 80% for novel objects. When the annotation budget is increased to 40 samples, the robot is able to grasp successfully more than 90% of the time for both seen and novel objects.**

I. Introduction

Grasping unknown objects in unstructured environments still represents a highly challenging task for robotic systems. Over the last years, data-driven deep learning methods have been proven to solve such tasks effectively and are being increasingly applied in real-world robotic systems for industry or logistics. However, in order to generalize to unknown target data domains, such passively trained grasping systems usually rely on large datasets, which are costly to collect and can result in limited zero-shot performance, especially when using synthetic data. This need for large training datasets is a significant bottleneck for current data-driven robotic systems.

In response to this training data challenge, our work proposes a sample-efficient learning framework for a grasping robot based on active learning. Building up on prior knowledge from simulation data, our proposed methodology enables cost-effective adaptation to the real-world target data domain, thereby improving grasp performance and deployment time. In our active learning framework, the robot perception system actively queries a pool of unlabeled data for the most informative data points from which to learn. This methodology assumes that some data samples provide more value to the training process than others, given the current state of the model. By focusing on these informative data samples, an active learning-based robotic system has the potential to quickly learn new tasks or adapt to new data domains in a more efficient and cost-effective manner.

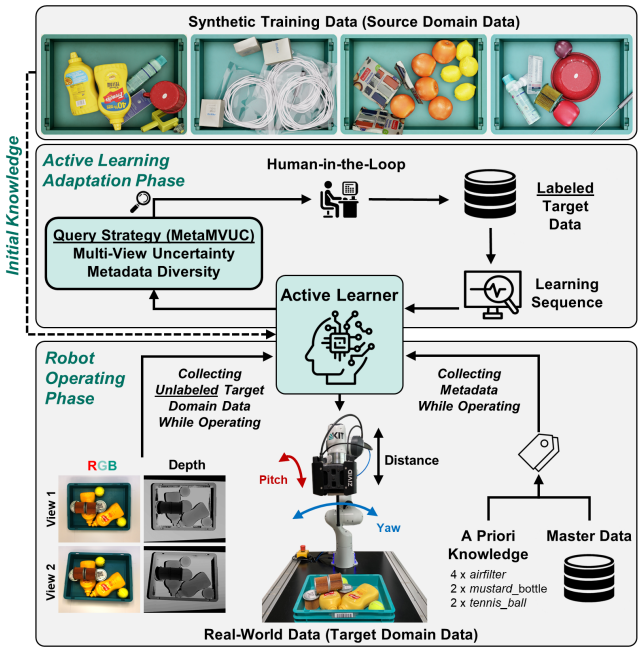


Fig. 1: Our established active learning framework for a grasping robot, enabling sample-efficient sim-to-real domain adaptation. Starting from synthetic data, the model parameters are incrementally adapted to the real-world target data domain. The proposed query strategy, **MetaMVUC**, iteratively queries a pool of unlabeled samples leveraging multi-view uncertainty and metadata diversity to select a set of informative and well distributed samples to learn from in each learning round.

In our work, we establish a novel active learning framework specifically designed for robotic grasping systems (cf. Fig. 1). Synthetic pre-training of robotic systems has shown promising capabilities for zero-shot transfer to the real world, providing a potentially cost-effective and robust initial knowledge base for active learning strategies. Robots equipped with arm-mounted cameras can actively sense their environment from multiple viewpoints. Often operating in metadata-rich domains such as agriculture, industry, or logistics, these robots are ideally suited to exploit active learning techniques based on advanced uncertainty or diversity measures. A pivotal element of every active learning system is the implemented query strategy. Our proposed query function, MetaMVUC, focuses on multi-view uncertainty and leverages metadata knowledge to select both informative and well-distributed data points to learn from. We base our approach on the assumption that inconsistencies in model predictions across different camera viewpoints serve as a reliable measure of model uncertainty. A concept which has been previously demonstrated in [34] for the task of semantic segmentation of indoor scenes. While exploiting model uncertainty to query informative samples is a common strategy for active learning, its use can result in poor data distribution coverage due to a lack of diversity consideration in the query strategy, particularly evident in early training episodes (cold start problem) [8]. To mitigate this issue, we employ synthetic pre-training as a 'warm start' strategy, enabling more effective deployment in real-world applications. Furthermore, we conclude that in many real-world robotic applications, metadata, including the classes and quantities of objects present in the grasp scene is often available, as it is needed for higher-level tasks. Inspired by the recent trend towards hybrid active learning methods that combine uncertainty and diversity measures, our proposed query strategy, MetaMVUC, shows how such metadata can improve the sample diversity of our active learning framework.

Our contributions can be summarized as follows:

- We design a novel, hybrid query strategy, MetaMVUC, which leverages our proposed multi-view uncertainty scoring combined with metadata diversity scoring.
- We establish an active learning framework for bin picking which enables fast model adaptation on real robots.
- We evaluate our methods on the MGNv2-Real dataset, as well as through physical grasping experiments in high clutter in our physical robot cell.

II. Related Work

A. Scene-Aware Robot Grasping

Reliable robotic grasping requires a vision system capable of detecting objects of interest, inferring suitable grasp poses, and reasoning about suitable object manipulation sequences.

Vacuum Grasp Detection: In automation and logistics, data-driven vacuum grasping has become a standard practice due to its versatility and effectiveness in handling common challenges such as flat bags, tightly packed items, and narrow bins. Besides their compact design, the rotational symmetry of vacuum suction cups simplifies grasp pose detection by reducing the degrees of freedom to four, presenting a simpler alternative to parallel-jaw gripper or multi-fingered hands, which often have seven or more degrees of freedom. Vacuum grasp detection is commonly framed either as a pixel-wise regression task of grasp quality heatmaps indicating suctionable areas for the whole image in a single shot [6, 23, 16] or two-stage sample-based approaches [27, 43, 44], where samples are ranked by their respective score [26].

Object Relationship Reasoning: Challenges in robotic grasping arise when items are arranged in clutter and overlap, leading to accidental simultaneous grasping of multiple items or failed attempts due to excessive contact forces [29]. Recent work addresses this object layout challenge by various means, including the detection of a full object relationship tree [40, 41, 42, 47, 9], amodal instance segmentation masks [2], or more simply by assessing the occlusion properties of individual objects without considering their adjacent relationships [16]. For goal-directed grasping of potentially occluded target objects (*singulation task*), it is crucial to understand the full relationship among objects based on amodal segmentation masks or relationship trees. However, research shown in [16] indicates that for the task of reliably emptying a cluttered scene of objects (*decluttering task*), focusing solely on the detection of occlusion properties of objects is sufficient.

Proposed approach: Our grasping pipeline is based on [16]. Object instances are detected together with their occlusion class. Vacuum grasp detection is framed as pixel-wise regression task of a grasp quality heatmap for the whole image.

B. Active Learning for Computer Vision

The primary objective of active learning is to optimize model performance while minimizing the costs associated with annotating training data [31]. Given the intensive data requirements of deep learning models, active learning has become crucial for reducing the labeling effort involved in the supervised training of such deep learning models. Over time, a variety of deep active learning methods have been developed which can be categorized based on the availability of the unlabeled data (pool-based vs. stream-based) [5], the presence of initial model knowledge (cold start vs. warm start) [24], or the type of query function employed [28]. Query functions are often categorized into uncertainty and hybrid query strategies, diversity-based methods, or meta-learning-based approaches. This review concentrates on pool-based active learning methods, commonly applied in robotics [37, 11] and autonomous driving [21] due to their straightforward implementation. For a extended review about stream-based active learning methods we refer to [5].

Diversity-Based Query Strategies: Diversity-based query strategies aim to select a subset of samples that best represent the entire data distribution. They are based on the assumption that a good data coverage in the selected subset effectively filters out redundant samples or irrelevant outliers. Prominent methods for diversity-based active learning work by either jointly selecting a subset of samples that best covers the data

in feature space [32, 1, 19], or by iteratively selecting the most representative samples [17, 37, 35].

Uncertainty-Based and Hybrid Query Strategies: Uncertainty-based query strategies aim to select a subset of samples where the model is most uncertain about. They assume that data samples where the model already shows high confidence contribute less to the training than samples where the model's prediction are uncertain. Defining metrics for uncertainty in deep neural network's prediction has been ongoing research for many years [15]. Within the context of active learning, uncertainty-based query strategies commonly employ methods based on information entropy [31, 33], Bayesian models [25, 22, 12, 13], ensembles [3], learning loss [38], and consistency [34, 10, 18, 14, 39]. Hybrid query strategies combine the idea of uncertainty- and diversity-based sampling in one method, aiming for both informative and at the same time representative samples to query [30, 36, 45].

Proposed Approach: Our hybrid query strategy uses multi-view uncertainty and metadata diversity scoring for sample-efficient learning. It is related to ViewAL [34], which applies multi-view uncertainty to active learning for semantic segmentation, and CALD [39], which uses consistency across original and augmented images for bounding box detection.

III. Method

The training of an occlusion-aware object detection network f_{OD} based on Mask-RCNN [20] architecture and a vacuum grasp detection network f_{SC} based on DeepLabv3 [7], both introduced in [16], with real-world data is expensive and time-consuming. While simulation-based data is inexpensive to generate, it often suffers from a significant sim-to-real domain gap, resulting in limited zero-shot performance. In our work, we tackle this issue via active domain adaptation. Our proposed active learning framework adapts f_{OD} and f_{SC}, both networks pretrained on source data domain (*synthetic data*), to a new target data domain (*real-world data*) by using a small amount of labeled target data. To select the most informative target data domain samples to learn from, we introduce a novel query strategy, **MetaMVUC**, which will be described in detail in the following subsections.

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 $Y_{\mathcal{LS}}$. Meanwhile, $\mathcal{T}$ represents the target data domain, which at the beginning only contains unlabeled data samples $X_{\mathcal{UT}}$. Throughout the active learning 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 $\{(X_{\mathcal{LT}}, Y_{\mathcal{LT}})\}$.

In an active domain adaptation setup, the learning algorithm initially has access to $\mathcal{S}$, containing samples $\{(X_{\mathcal{LS}}, Y_{\mathcal{LS}})\}$. Following the pool-based active learning approach, the static pool of unlabeled data samples $X_{\mathcal{UT}}$ is available from the start. In total, the learner has a query budget B which is much smaller than the amount of unlabeled data $B \ll |X_{\mathcal{UT}}|$. For each learning round R_n, $n = 1 \dots N$, and with a per-round query budget of B/N, the learner queries a subset of samples from the pool of $X_{\mathcal{UT}}$ and requests the user to annotate them $X_{\mathcal{UT}} \rightarrow (X_{\mathcal{LT}}, Y_{\mathcal{LT}})$. Consequently, in each learning round R_n, the learner gains access to an by B/N incremented set of annotated target samples $X_{\mathcal{LT}}$, while the pool of target samples $X_{\mathcal{T}} = X_{\mathcal{LT}} \cup X_{\mathcal{UT}}$ remains static.

The goal of the learning algorithm is to optimize performance in object detection and grasp point detection within the new target data domain, while keeping the number of annotated samples $|X_{\mathcal{LT}}|$ small.

B. Multi-View Uncertainty Scoring

To obtain a good estimate of the uncertainty in the network prediction, we exploit the active vision capabilities of robotic arms equipped with hand-mounted cameras. Our pipeline for multi-view consistency-based uncertainty scoring (MVUC) is divided into the following steps, which will be described in more detail below (cf. Fig. 2): For a given object scene, the robot arm moves to two different camera viewpoints and captures sensor images. Network inference on both viewpoint-specific data samples is performed to predict vacuum grasp heatmaps and objects together with their occlusion properties. Reasoning about underlying network uncertainty is done by matching network predictions across both viewpoints.

Step 1: Multi-View Data Acquisition

Each object scene is observed from two different camera viewpoints C_i, $i = 1, 2$, defined as transformation matrix $\mathbf{T}_{C_i}^{W}$ relative to a chosen world coordinate system W. The first camera viewpoint $\mathbf{T}_{C_1}^{W}$ is positioned overhead, and the second viewpoint $\mathbf{T}_{C_2}^{W}$ is randomly selected from pre-configured viewpoints, varying in distance, pitch, and yaw angles. At each camera viewpoint C_i, a color image X_{RGB}^{i} and a depth image X_{D}^{i} are captured, stored along with the corresponding pose of the camera $\mathbf{T}_{C_i}^{W}$.

Step 2: Network Inference

For each camera viewpoint C_i, $i = 1, 2$, object instances together with their occlusion classes are detected given the color image X_{RGB}^{i} of the scene. Following the SSMP methodology proposed in [16], the object detection network f_{OD} detects object instances $O_l^i = f_{\mathrm{OD}}(X_{\mathrm{RGB}}^{i})$ pixel-wise along with their occlusion class: unoccluded, less than 10% occlusion, and greater than 10% occlusion. Additionally, a vacuum grasp heatmap V^i is predicted for each viewpoint using the depth image X_{D}^{i} of the scene. Following the SSMP grasp approach from [16], we interpret vacuum grasp detection as a pixel-wise vacuum graspability learning problem. The vacuum grasp heatmap V is obtained as a pixel-wise classification task across 25 bins, representing score values in the range of $[0, 1]$:

$$V^i = \operatorname{argmax}(\hat{V}^i) = \operatorname{argmax}\left(\operatorname{softmax}\left(f_{\mathrm{SC}}(X_{\mathrm{D}}^{i})\right)\right). \quad (1)$$

For more details, we refer to [16].

Step 3: Multi-View Uncertainty

In order to calculate the uncertainty based on multi-view predictions, it's necessary to match corresponding pixels across camera viewpoints C_i, $i = 1, 2$. We are interested in knowing which pixels in image X^1 and in X^2 cross-project to the

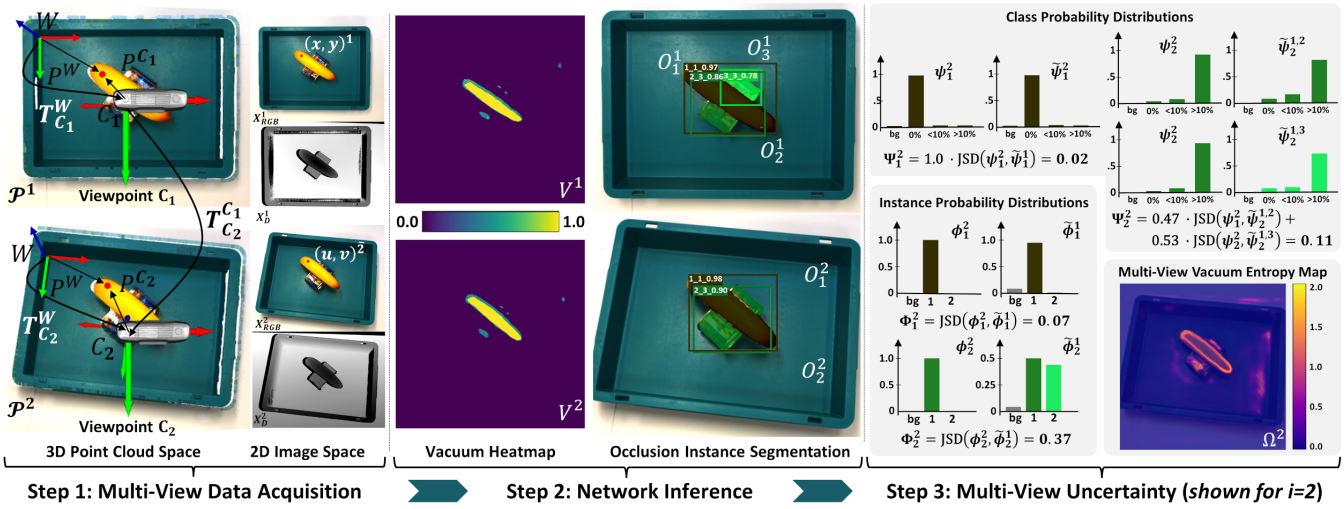


Fig. 2: Overview of the proposed multi-view consistency-based uncertainty scoring (MVUC) pipeline, divided into three subsequent steps: 1. Multi-view data acquisition using arm-mounted camera, 2. Network inference for both vacuum grasp detection and occlusion-aware object detection, and 3. Multi-view uncertainty estimation.

same real-world point P^W (cf. Fig. 2). Using the depth image X^i_{D} and the intrinsic camera matrix, we can project all pixels from image X^1 and X^2 into 3D space. As a result, we obtain two point clouds $\mathcal{P}^i = \{(x_k, y_k, z_k)^{C_i} \mid k = 1,\dots,K\}$, each containing points relative to the respective camera coordinate system C_1 and C_2. Given the poses of the camera viewpoints $\mathbf{T}^W_{C_1}$ and $\mathbf{T}^W_{C_2}$, all points from $\mathcal{P}^1$ and $\mathcal{P}^2$ can be projected into the common world coordinate system W. By aligning the points from both viewpoints into a common coordinate system W, it is possible to match points in $\mathcal{P}^1$ and $\mathcal{P}^2$ based on their relative distance in 3D space (ICP algorithm in [46]). Knowing the 2D pixel coordinates of registered points in 3D space, we achieve pixel-wise cross-projection across images X^1 and X^2, captured at camera viewpoints C_1 and C_2.

For the proposed multi-view uncertainty scoring, we are interested in knowing how consistently the network predicts across viewpoints. When it comes to vacuum grasping, inspired by [34], given a reference viewpoint i, the mean softmax distribution of cross-projected vacuum predictions $\hat{V}$ is computed. We calculate the pixel-wise entropy H of the mean softmax, and obtain the multi-view vacuum entropy map Ω^i as:

$$\Omega^i(u,v) = H\left(\frac{\hat{V}^i(u,v) + \hat{V}^j(x,y)}{2} \,\middle|\, \begin{array}{l} X^j_{\mathrm{D}}(x,y)\,\mathrm{cross}- \\ \mathrm{projects}\, X^i_{\mathrm{D}}(u,v) \end{array}\right), \tag{2}$$

where $j = 3 - i$ for $i = 1, 2$. Here, $(x, y)^j$ stands for a cross-projected pixel from viewpoint j, given a pixel $(u, v)^i$ in the reference viewpoint i (cf. Fig. 2 for $i = 2$). Ω^i is computed for both viewpoints $i = 1, 2$ and we obtain as a result the **multi-view vacuum entropy (MVVH)** score as:

$$\mathrm{MVVH} = \frac{1}{2}\left(\frac{1}{|\Omega^1|}\sum_{v=1}^{H}\sum_{u=1}^{W}\Omega^1 + \frac{1}{|\Omega^2|}\sum_{v=1}^{H}\sum_{u=1}^{W}\Omega^2\right), \tag{3}$$

where W and H are the width and height of the image X^i_{D}.

Besides evaluating MVVH, we propose a metric to assess the consistency of semantic instance predictions across viewpoints, which is crucial for occlusion-aware instance segmentation or other instance segmentation tasks. Unlike semantic segmentation, which classifies pixels into semantic classes independently, instance segmentation additionally groups pixels into coherent sets (instances) that share a common identifier. Our method focuses on computing the consistency of instance predictions across viewpoints. We assume that the network is confident about an object if it is detected coherently from both viewpoints.

To quantify the instance consistency Φ^i_l for a detected object instance O^i_l in a reference viewpoint i, the Jensen-Shannon divergence (JSD) is utilized. It measures the similarity between corresponding instance probability distributions ϕ^i_l and $\tilde{\phi}^j_l$:

$$\Phi^i_l = \mathrm{JSD}\left(\phi^i_l, \tilde{\phi}^j_l\right), \tag{4}$$

where $j = 3 - i$ for $i = 1, 2$. Here, $\tilde{\phi}^j_l$ characterizes the distribution of corresponding instance predictions in viewpoint j, given a detection O^i_l in the reference viewpoint i. To compute $\tilde{\phi}^j_l$, we adopt a pixel-wise approach. Specifically, we count the number of cross-projected pixels associated with each instance prediction in viewpoint j, given the detection O^i_l in the reference viewpoint i. In cases where cross-projected pixels have no detection in viewpoint j, these pixels are categorized as background pixels. The resulting counts are normalized by the total sum of the cross-projected pixels and sorted in descending order, yielding a probability distribution $\tilde{\phi}^j_l$ for viewpoint j (cf. Fig. 2, illustrated for $i = 2$ and $j = 1$). Φ^i_l is computed for all detected object instances O^i_l in both viewpoints $i = 1, 2$ and we obtain as a result the **multi-view instance consistency (MVIC)** score as:

$$\mathrm{MVIC} = \frac{1}{2}\left(\frac{1}{N}\sum_{l=1}^{N}\Phi^1_l + \frac{1}{M}\sum_{l=1}^{M}\Phi^2_l\right), \tag{5}$$

where N and M are the total number of detected object instances in image X^1_{RGB} and X^2_{RGB}, respectively.

Our methodology further extends to quantify the multi-view consistency of class probability distributions for detected object instances across viewpoints. Given a reference viewpoint i, the multi-view class consistency Ψ_l^i for a detected object instance O_l^i in the reference viewpoint i is calculated using the Jensen-Shannon divergence. It measures the similarity between corresponding class probability distributions ψ_l:

$$\Psi_l^i = \mathrm{JSD}\left(\psi_l^i, \tilde{\psi}_l^j\right), \tag{6}$$

where $j = 3 - i$ for $i = 1, 2$. Here, $\tilde{\psi}_l^j$ characterizes the class probability distributions of corresponding object instance predictions in the second viewpoint j, given a detection O_l^i with a class probability distribution ψ_l^i in the reference viewpoint i. To obtain Ψ_l^i, all detections in viewpoint j corresponding to the reference detection O_l^i are weighted according to their pixel-surface area compared to the total sum of cross-projected pixels. Ψ_l^i is computed for all detected object instances O_l^i in both viewpoints $i = 1, 2$ and we obtain the **multi-view class consistency (MVCC)** score as:

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

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

These proposed metrics – MVVH, MVIC, and MVCC – serve as robust tools for evaluating the consistency and uncertainty of vacuum grasps and semantic instance detections using a robot's multi-view sensing capabilities. In our active learning framework, for each query round, we apply these metrics to all samples in $X_{\mathcal{UT}}$. The results are sorted in descending order, yielding individual rankings ν_Ω, ν_Φ, and ν_Ψ for MVVH, MVIC, and MVCC, respectively.

C. Metadata Diversity Scoring

Research has shown that solely using uncertainty-based query functions often results in redundant sample selection [32]. To address this, diversity-based query methods (cf. Sec. II-B) often aim to identify the most representative samples in a data distribution. This is often done by transforming samples into high-dimensional feature vectors and applying distance metrics for clustering [32, 30]. However, these approaches frequently encounter the curse of dimensionality, where increased dimensions tend to make distance metrics less informative.

In our work, we use metadata for diversity scoring, taking into account its widespread availability in many logistics and industrial applications, as such data is often required to perform higher-level tasks. We propose a scoring algorithm which makes use of such information to effectively query diverse and balanced set of samples. Specifically, we assume prior knowledge of each object's semantic class, the number of class instances in a scene, and the primary material of an object. Our proposed algorithm iterates through all unlabeled samples $X_{\mathcal{UT},j}$, $j = 1 \ldots |X_{\mathcal{UT}}|$ and ranks the contribution of each sample $X_{\mathcal{UT},j}$ in terms of semantic class distribution ρ_C, number of instances per scene distribution ρ_n, and instance material distribution ρ_M to a uniform distribution $\tilde{\rho}_C$, $\tilde{\rho}_n$, and $\tilde{\rho}_M$. In order to measure the contribution of a data sample $X_{\mathcal{UT},j}$, the Jason-Shennon divergence is computed between the set of labeled samples plus the considered sample $\{X_{\mathcal{LT}}, X_{\mathcal{UT},j}\}$ and the uniform distributions $\tilde{\rho}_C$, $\tilde{\rho}_n$, and $\tilde{\rho}_M$. It is computed for all three metadata attributes and we obtain the **metadata diversity score (MDDS)** as:

$$MDDS = \frac{1}{3}\left(\mathrm{JSD}\left(\rho_C(X_{\mathcal{LT}} + X_{\mathcal{UT},j}), \tilde{\rho}_C\right) + \right. \tag{8}$$

$$\mathrm{JSD}\left(\rho_n(X_{\mathcal{LT}} + X_{\mathcal{UT},j}), \tilde{\rho}_n\right) + \tag{9}$$

$$\left.\mathrm{JSD}\left(\rho_M(X_{\mathcal{LT}} + X_{\mathcal{UT},j}), \tilde{\rho}_M\right)\right). \tag{10}$$

In our active learning framework, MDDS is applied to all samples in $X_{\mathcal{UT}}$, and the resulting scores are ranked in ascending order to obtain a metadata diversity ranking ν_ζ.

D. Hybrid Query Strategy

To take advantage of both diversity-based and uncertainty-based query strategies, we integrate the proposed multi-view uncertainty (cf. Sec. III-B) and metadata diversity scoring metrics (cf. Sec. III-C) into a hybrid query strategy entitled **MetaMVUC**. Based on the previously discussed metrics MVVH, MVIC, MVCC, and MDDS, we obtain individual sample rankings ν_Ω, ν_Φ, ν_Ψ, and ν_ζ for each metric, respectively. In order to achieve an aggregated ranking across all samples, we employ the Borda rule method, a well-established positional voting rule, where each metric MVVH, MVIC, MVCC, and MDDS acts as voter $v \in \mathbb{V} = \{\Omega, \Phi, \Psi, \zeta\}$. Following [4], the rank of the j-th sample, $X_{\mathcal{UT},j}$, is denoted as $\mathrm{rk}\left(X_{\mathcal{UT},j}, \nu_v\right)$ under the voter v. For example, $\mathrm{rk}\left(X_{\mathcal{UT},j^*}, \nu_\Omega\right) = 1$ expresses that the sample j^* is the highest-ranked sample in ν_Ω according to the MVVH metric. The Borda rule assigns each sample in $X_{\mathcal{UT}}$ a score based on its aggregated ranking. Specifically, the Borda score for each sample $X_{\mathcal{UT},j} \in X_{\mathcal{UT}}$ is calculated as the λ_v weighted inverse of its voter-specific rank $|X_{\mathcal{UT}}| - \mathrm{rk}\left(\mathrm{X}_{\mathcal{UT},\mathrm{j}}, \nu_\mathrm{v}\right)$, and summed over all voters $v \in \mathbb{V}$:

$$\mathrm{Borda}\left(X_{\mathcal{UT},j}\right) = \sum_{v \in \mathbb{V}} \lambda_v \cdot \left(|X_{\mathcal{UT}}| - \mathrm{rk}\left(\mathrm{X}_{\mathcal{UT},\mathrm{j}}, \nu_\mathrm{v}\right)\right), \tag{11}$$

with the weights chosen as: $\lambda_\Phi = 2$ and $\lambda_\Omega = \lambda_\Psi = \lambda_\zeta = 1$.

This scoring system is designed to identify samples that achieve broad consensus across all metrics, aiming to form informative and well-distributed query sets. Given our per-round query budget $b_n = B/N$, we select the top-b highest-ranked samples based on their Borda scores and submit them for annotation. Once annotated, these samples are added to the training dataset $\{(X_{\mathcal{LT}}, Y_{\mathcal{LT}})\}$ and used for iterative model adaptation to the real-world target domain $\mathcal{T}$ (cf. Fig. 1).

IV. Experiments

In our experiments, we assess the efficiency of our proposed hybrid active query strategy, **MetaMVUC**, within a real-world scenario of robotic vacuum grasping in cluttered environments. The central research question we address is as follows: How does **MetaMVUC** compare to a baseline random sampling

(a) (b)

Fig. 3: (a) Real-world robot cell used for evaluating our proposed active learning framework and query strategy, MetaMVUC. It is equipped with a Zivid Two RGB-D camera and a Festo vacuum gripper system featuring a 20 mm suction cup. (b) Overview of two objects sets used for evaluation: seen objects and unseen objects

approach (**RAND**) and the widely-applied active learning method, coreset [32] (**CORE**), in terms of enhancing occlusion-aware object detection and grasp performance across active learning rounds R_n and number of annotated samples $|X_{\mathcal{LT}}|$?

A. Experimental Setup

Dataset and Real-World Setup: We perform experiments on the real-world split of MetaGraspNetv2 dataset [16] (MGNv2-Real) and physical grasp experiments in our real-world robot cell (cf. Fig. 3). In addition to its comprehensive coverage and focus on task-specific challenges for bin picking, the MetaGraspNetV2 dataset also offers practical advantages. Since the dataset was collected using a setup identical to that employed in our active learning experiments, we can replicate the data acquisition settings. This allows us to mimic a realistic pool-based active learning environment in the real world, without the need for extensive labelling, as the data comes already pre-annotated.

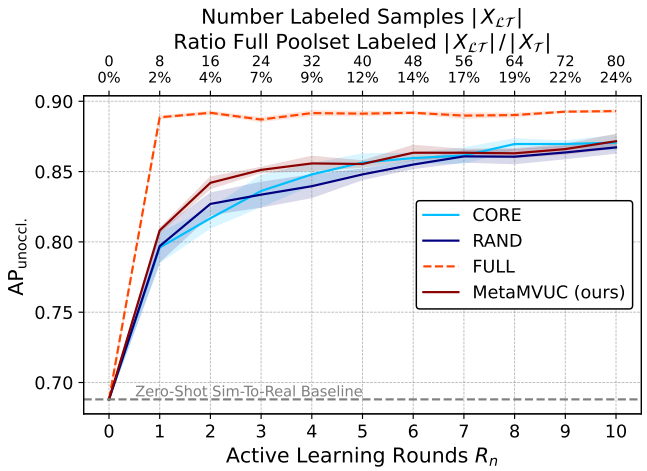
Experimental Design For our real-world experiments, we select 33 objects from the MGNv2 dataset, referred to as *seen* objects, and introduce 23 novel objects, referred to as the *unseen* objects (cf. Fig. 3b). We filter the MGNv2 dataset for scenes containing only objects from the seen object set. The remaining 233 scenes are each annotated for an additional second viewpoint and then divided into two sets: 70% form the pool set (MGNv2-Pool), and the remaining 30% form the test set (MGNv2-Test). This experimental design allows us to evaluate our methods in a traditional pool-based active learning setting (MGNv2-Test + seen objects), while also introducing a more challenging generalizability task. Specifically, this task involves grasping novel (unseen) objects that are not included in the pool set MGNv2-Pool. Although this task is less discussed in the active learning literature, it has great practical relevance. It effectively represents real-world conditions where data distributions can shift and robots have to deal with novel objects.

Training Details: As starting point, we use network weights for f^{sc} and f^{obj} pre-trained on the synthetic dataset MGNv2-Sim [16]. In our active learning experiments, we use Adam and SGD optimizer for training f^{obj} and f^{sc}, respectively, with learning rates set to 0.001 and $1e-5$. A batch size of 4 is chosen and the per-round query budget is set to $b_n = 4$, chosen intentionally small to reflect the high cost of annotating cluttered scenes in the real-world. Each queried scene is annotated for both viewpoints, resulting in a total of 8 annotated samples per query round. We run our experiments for $N = 10$ query rounds R_n, $n = 1 \dots N$, and perform 20 training epochs per round R_n (full iteration through queried training set). To mitigate stochastic effects, each experiment is repeated 5 times using alternating random seeds. The results reported are the mean averages of these trials, unless specified otherwise.

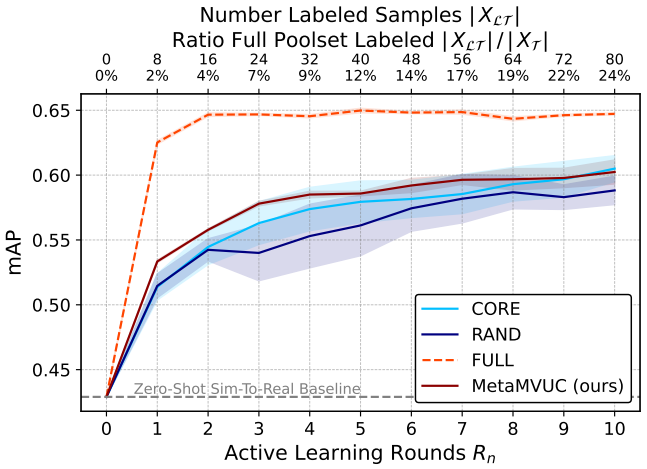
Metrics: For evaluating occlusion-aware object detection, two metrics are employed: the Average Precision for objects in the unoccluded class ($\text{AP}_{\text{unoccl.}}$), calculated at IoU thresholds ranging from 0.5 to 0.95 in increments of 0.05, and the mean Average Precision (mAP) across all three occlusion classes. In our real-world grasp experiments, the performance is measured using the following metrics: the number of successful grasps over the total number of 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).

B. Results on MGNv2-Test

Using MGNv2-Pool as the pool set to query from, we evaluate the performance of MetaMVUC on the MGNv2-Test dataset for occlusion-aware object detection. Results are reported for $\text{AP}_{\text{unoccl.}}$ (cf. Fig. 4a) and mAP (cf. Fig. 4b) for different sizes of queried sets $|X_{\mathcal{LT}}|$ and learning rounds R_n. As illustrated in Fig. 4, the results show that MetaMVUC outperforms a state-of-the-art active learning method CORE, and a random sampling baseline RAND in terms of $\text{AP}_{\text{unoccl.}}$ and mAP, especially in early active learning rounds and smaller query set sizes (cf. Fig. 4a $R_n, n \leq 4$ and cf. Fig. 4b $R_n, n \leq 7$). This demonstrates that MetaMVUC is able to adapt fast and sample-efficient to target data domain distributions, given a small annotation budget. As the learning progresses, MetaMVUC and CORE show competitive performance. In general, our experiments show that the performance difference, especially in terms of $\text{AP}_{\text{unoccl.}}$, between active sampling methods (CORE and MetaMVUC) and random sampling RAND decreases as the number of annotated samples increases. For comparison, we also include the FULL scenario, in which the learner was given full access to the entire annotated pool set, serving as the upper baseline. As expected, the use of



(a) $\mathrm{AP}_{\mathrm{unoccl.}}$ for MGNv2-Test Split



(b) mAP for MGNv2-Test Split

Fig. 4: Results for occlusion-aware object detection.

the entire annotated pool set for learning results in superior performance. However, it is noteworthy that in early rounds, MetaMVUC achieves respectable results compared to FULL, despite using only a fraction of the training data FULL was trained on (cf. Fig. 4).

C. Results in Real-World Cell

In our real-world robot experiments, we investigate the grasping performance of a robot trained using our proposed active learning method, MetaMVUC. Specifically, we aim to determine how MetaMVUC influences sample efficiency and grasp performance in terms of successful grasp rate R_{grasp} and autonomously cleared object rate R_{obj}. To address this, we evaluate different model checkpoints for f_{OD} and f_{SC} from active learning rounds R_2, R_5, and R_{10}, each queried by MetaMVUC, by conducting real-world grasping experiments in our robotic cell (cf. Fig. 3a). For comparative analysis, we also perform experiments without MetaMVUC (R_0), representing a zero-shot sim-to-real baseline. Our grasping pipeline is similar to the proposed SSMP algorithm in [16], which ranks grasp proposals from f_{SC} based on object detections from f_{OD} together with their occlusion predictions. However, in contrast to the method proposed in [16], our approach for grasping captures a new image after each grasp attempt.

We perform experiments on two object sets (cf. Fig. 3b): seen objects, also part of the pool set, and unseen objects, not part of the pool set and therefore novel at test time. At the beginning of each run, the grasp scene consists of 10 objects, arranged in high clutter (cf. Fig. 3a). The robot is tasked to empty the whole scene. A grasp is considered successful if the object has been picked up and transferred to another bin. After two failed grasp attempts per object and run, the object is removed manually by a human supervisor. After each active learning round $R_n, n = 0, 2, 5, 10$, the robot attempts to grasp a total of 100 objects, distributed across 10 runs (5 runs of seen objects and 5 runs of unseen objects). Reported numbers in Table I and Fig. 5 are averaged across all runs.

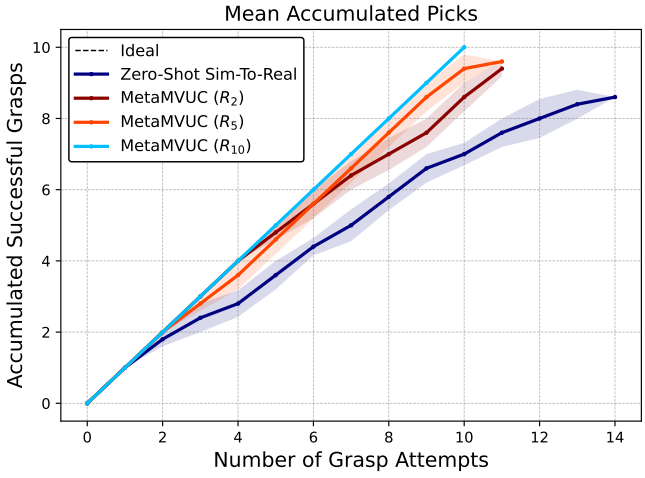
The results in Table I show, that even in early rounds of active learning, remarkable real-world performance can be achieved by MetaMVUC in terms of successful grasp rate R_{grasp} and autonomously cleared object rate R_{obj}. After 5 learning rounds R_5, which is equivalent to 40 annotated images out of a pool set size of 324, a grasping robot trained with MetaMVUC achieves more than 90% successful grasp rate R_{grasp} and autonomously cleared object rate R_{obj} for both seen and unseen objects. Furthermore, a direct comparison between zero-shot sim-to-real and early rounds of active learning with MetaMVUC (cf. Fig. 5 and Table I for R_2) shows a strong performance boost in terms of R_{grasp} and R_{obj} of at least 20% and 8%, respectively, even with only 16 out 324 annotated samples for R_2.

For seen objects (cf. Fig. 5a and cf. Table I), grasping performance increases as the size of the training dataset increases. This is an expected result, given that the test objects considered in this experiment are part of the training dataset. Remarkable, after 10 runs of active learning, our grasping robot (light blue line for MetaMVUC (R_{10})) achieves performance that matches the ideal (cf. Fig. 5a). For unseen objects, we can observe that performance is best after 5 learning rounds and then drops slightly again (cf. Fig. 5b and cf. Table I). A possible explanation could be that the model tends to overfit to the training data distribution as the number of training epochs increases. This represents a general problem of static, pool-based active learning approaches, as the data pool from which queries are made remains static by definition.

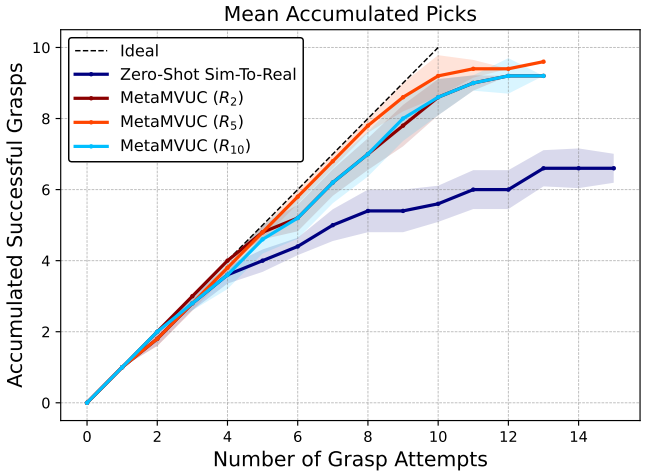
Nevertheless, even for the challenging task of unseen data at test time, based on our experimental results, it can be summarized that our established active learning framework built upon our proposed query stratey, MetaMVUC, is an effective and robust approach for active domain adaptation of grasping robots, even when the annotation budget is small.

V. Conclusion

In this paper, an active learning framework for robot learning has been designed, enabling sample-efficient training of real-world grasping robots. The proposed framework uses synthetic data as an initial starting point and then employs a novel,



(a) Seen Object Set



(b) Unseen Object Set

Fig. 5: Results are presented from physical robot grasping experiments in highly cluttered environments. The average number of autonomously picked items, for both seen objects in Fig. 5a and unseen objects in Fig. 5b, is plotted against the total number of grasp attempts. The dashed line represents the ideal scenario where all items are picked on the first attempt. Top left means better performance, characterized by a high number of autonomously picked objects in combination with a low number of grasp attempts. The standard deviation, scaled by a factor of 0.5, is shown in the shaded areas.

hybrid query strategy, MetaMVUC, to identify the most relevant samples to learn from. Our proposed method, MetaMVUC, uses multi-view uncertainty and metadata diversity scoring in order to find the samples that are both highly informative for the learner and at the same well representative of the overall data distribution. Experiments on the MGNv2 dataset and in our real-world robot cell have demonstrated its effectiveness and robustness. Our method significantly reduces the number of annotated samples required, effectively generalizes to unseen objects, and increases both the successful grasp rate and the autonomously cleared object rate, achieving improvements of at least 20% and 8%, respectively, over the zero-shot sim-to-real baseline. Moreover, real-world grasp experiments demonstrate that with just 16 annotated data samples selected out of 324 pool set samples, our system achieves successful grasp rates of over 87% for seen objects and 80% for novel objects. When the annotation budget is increased to 40 samples, the robot grasps successfully more than 90% of the time for both seen and novel objects. Given the high costs associated with data collection and annotation, sample-efficient robot learning systems are of great importance. Our proposed learning framework and query strategy, MetaMVUC, can contribute to the development of low-cost and rapidly deployable robots for real-world grasping.

TABLE I: Real-world vacuum grasp performance R_{grasp} and R_{obj} with standard deviation in parentheses across active learning rounds R for seen and unseen objects in clutter.

Method	Rounds R	R_{grasp} (%)		R_{obj} (%)	
		Seen Objects	Unseen Objects	Seen Objects	Unseen Objects
Zero-Shot S2R[a]	0	67.6 (8.9)	46.2 (8.7)	86.0 (8.0)	66.0 (10.2)
MetaMVUC (ours)	2	87.3 (7.3)	80.2 (10.7)	94.0 (4.9)	92.0 (4.0)
	5	92.7 (8.9)	90.2 (12.6)	96.0 (4.9)	96.0 (4.9)
	10	100.0 (0.0)	81.7 (12.2)	100.0 (0.0)	92.0 (7.5)

[a] S2R: Sim-To-Real

ACKNOWLEDGMENTS

R. Rayyes's position is funded by the Baden-Wurttemberg Ministry of Science, Research and the Arts within Innovations Campus Mobilität der Zukunft (ICM).

REFERENCES

[1] Sharat Agarwal, Himanshu Arora, Saket Anand, and Chetan Arora. Contextual diversity for active learning. In *Computer Vision – ECCV 2020*, pages 137–153, 2020. doi: 10.1007/978-3-030-58517-4_9. URL https://link.springer.com/chapter/10.1007/978-3-030-58517-4_9.

[2] Seunghyeok Back, Joosoon Lee, Taewon Kim, Sangjun Noh, Raeyoung Kang, Seongho Bak, and Kyoobin Lee. Unseen object amodal instance segmentation via hierarchical occlusion modeling. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5085–5092, 2022. doi: 10.1109/ICRA46639.2022.9811646. URL https://ieeexplore.ieee.org/document/9811646.

[3] William H. Beluch, Tim Genewein, Andreas Nurnberger, and Jan M. Kohler. The power of ensembles for active learning in image classification. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9368–9377, 2018. doi: 10.1109/CVPR.2018.00976. URL https://ieeexplore.ieee.org/document/8579074/.

[4] Dávid Burka, Clemens Puppe, László Szepesváry, and Attila Tasnádi. Voting: A machine learning approach. *European Journal of Operational Research*, 299(3):1003–1017, 2022. doi: 10.1016/j.ejor.2021.10.005. URL https://www.sciencedirect.com/science/article/pii/S037722172100850X.

[5] Davide Cacciarelli and Murat Kulahci. Active learning for data streams: a survey. *Machine Learning*, 113(1): 185–239, 2024. doi: 10.1007/s10994-023-06454-2. URL https://link.springer.com/10.1007/s10994-023-06454-2.

[6] Hanwen Cao, Hao-Shu Fang, Wenhai Liu, and Cewu Lu. SuctionNet-1billion: A large-scale benchmark for suction grasping. *IEEE Robotics and Automation Letters*, 6(4): 8718–8725, 2021. doi: 10.1109/LRA.2021.3115406. URL https://ieeexplore.ieee.org/abstract/document/9547830.

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. URL https://arxiv.org/abs/1706.05587.

[8] Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu, Bihan Wen, Alan L. Yuille, and Zongwei Zhou. Making your first choice: To address cold start problem in vision active learning. *arXiv preprint arXiv:2210.02442*, 2022. URL https://arxiv.org/abs/2210.02442.

[9] Mengyuan Ding, Yaxin Liu, Chenjie Yang, and Xuguang Lan. Visual manipulation relationship detection based on gated graph neural network for robotic grasping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1404–1410, 2022. doi: 10.1109/IROS47612.2022.9981077. URL https://ieeexplore.ieee.org/document/9981077.

[10] Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixé, and Jose M. Alvarez. Not all labels are equal: Rationalizing the labeling costs for training object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14472–14481, 2022. doi: 10.1109/CVPR52688.2022.01409. URL https://ieeexplore.ieee.org/abstract/document/9878602.

[11] Jianxiang Feng, Jongseok Lee, Maximilian Durner, and Rudolph Triebel. Bayesian active learning for sim-to-real robotic perception. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10820–10827, 2022. doi: 10.1109/IROS47612.2022.9982175. URL https://ieeexplore.ieee.org/abstract/document/9982175.

[12] Yidan Feng, Biqi Yang, Xianzhi Li, Chi-Wing Fu, Rui Cao, Kai Chen, Qi Dou, Mingqiang Wei, Yun-Hui Liu, and Pheng-Ann Heng. Towards robust part-aware instance segmentation for industrial bin picking. In *International Conference on Robotics and Automation (ICRA)*, pages 405–411, 2022. doi: 10.1109/ICRA46639.2022.9811728. URL https://ieeexplore.ieee.org/document/9811728.

[13] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning (ICML)*, page 1183–1192, 2017. URL https://proceedings.mlr.press/v70/gal17a/gal17a.pdf.

[14] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö. Arık, Larry S. Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *Computer Vision – ECCV 2020*, pages 510–526, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123550511.pdf.

[15] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56 (1):1513–1589, 2023. doi: 10.1007/s10462-023-10562-9. URL https://doi.org/10.1007/s10462-023-10562-9.

[16] Maximilian Gilles, Yuhao Chen, Emily Zhixuan Zeng, Yifan Wu, Kai Furmans, Alexander Wong, and Rania Rayyes. MetaGraspNetV2: All-in-one dataset enabling fast and reliable robotic bin picking via object relationship reasoning and dexterous grasping. *IEEE Transactions on Automation Science and Engineering*, pages 1–19, 2023. doi: 10.1109/TASE.2023.3328964. URL https://ieeexplore.ieee.org/document/10309974.

[17] Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning. *arXiv preprint arXiv:1907.06347*, 2019. URL https://arxiv.org/abs/1907.06347.

[18] S. Alireza Golestaneh and Kris M. Kitani. Importance of self-consistency in active learning for semantic segmentation. In *British Machine Vision Conference (BMVC)*, 2020. URL https://arxiv.org/pdf/2008.01860.pdf.

[19] Chengcheng Guo, Bo Zhao, and Yanbing Bai. DeepCore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pages 181–195, 2022. doi: 10.1007/978-3-031-12423-5_14. URL https://link.springer.com/chapter/10.1007/978-3-031-12423-5_14.

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *arXiv preprint arXiv:1703.06870*, 2017. URL https://arxiv.org/abs/1703.06870.

[21] Aral Hekimoglu, Adrian Brucker, Alper Kagan Kayali, Michael Schmidt, and Alvaro Marcos-Ramiro. Active learning for object detection with non-redundant informative sampling. *arXiv preprint arXiv:2307.08414*, 2023. URL https://arxiv.org/abs/2307.08414.

[22] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011. URL http://arxiv.org/abs/1112.5745.

[23] Ping Jiang, Junji Oaki, Yoshiyuki Ishihara, Junichiro Ooga, Haifeng Han, Atsushi Sugahara, Seiji Tokura, Haruna Eto, Kazuma Komoda, and Akihito Ogawa. Learning suction graspability considering grasp quality and robot reachability for bin-picking. *Frontiers in Neurorobotics*, 16, 2022. doi: 10.3389/fnbot.2022.806898. URL https://www.frontiersin.org/articles/10.3389/fnbot.2022.806898/full.

[24] Qiuye Jin, Mingzhi Yuan, Shiman Li, Haoran Wang, Manning Wang, and Zhijian Song. Cold-start active learning for image classification. *Information Sciences*, 616:16–36, 2022. doi: 10.1016/j.ins.2022.10.066. URL https://linkinghub.elsevier.com/retrieve/pii/

S0020025522011768.

[25] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: efficient and diverse batch acquisition for deep bayesian active learning. In *International Conference on Neural Information Processing Systems (NeurIPS)*, pages 7026–7037, 2019. URL https://dl.acm.org/doi/pdf/10.5555/3454287.3454918.

[26] Kilian Kleeberger, Richard Bormann, Werner Kraus, and Marco F Huber. A survey on learning-based robotic grasping. *Current Robotics Reports*, 1:239–249, 2020. doi: 10.1007/s43154-020-00021-6. URL https://link.springer.com/article/10.1007/s43154-020-00021-6.

[27] Jeffrey Mahler, Matthew Matl, Xinyu Liu, Albert Li, David Gealy, and Ken Goldberg. Dex-net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5620–5627, 2018. doi: 10.1109/ICRA.2018.8460887. URL https://ieeexplore.ieee.org/abstract/document/8460887.

[28] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023. doi: 10.1007/s10462-022-10246-w. URL https://link.springer.com/article/10.1007/s10462-022-10246-w.

[29] Zherong Pan, Andy Zeng, Yunzhu Li, Jingjin Yu, and Kris Hauser. Algorithms and systems for manipulating multiple objects. *IEEE Transactions on Robotics*, 39 (1):2–20, 2023. doi: 10.1109/TRO.2022.3197013. URL https://ieeexplore.ieee.org/abstract/document/9893496.

[30] Viraj Prabhu, Arjun Chandrasekaran, Kate Saenko, and Judy Hoffman. Active domain adaptation via clustering uncertainty-weighted embeddings. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8485–8494, 2021. doi: 10.1109/ICCV48922.2021.00839. URL https://ieeexplore.ieee.org/document/9710350.

[31] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021. doi: 10.1145/3472291. URL https://dl.acm.org/doi/abs/10.1145/3472291.

[32] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/pdf?id=H1aIuk-RW.

[33] Burr Settles. Active learning literature survey. *CS Technical Reports*, 2009. URL https://minds.wisconsin.edu/handle/1793/60660.

[34] Yawar Siddiqui, Julien Valentin, and Matthias Niessner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9430–9440, 2020. doi: 10.1109/CVPR42600.2020.00945. URL https://ieeexplore.ieee.org/document/9156651.

[35] Samrath Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5971–5980, 2019. doi: 10.1109/ICCV.2019.00607. URL https://ieeexplore.ieee.org/abstract/document/9009538.

[36] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. Active adversarial domain adaptation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 728–737, 2020. doi: 10.1109/WACV45572.2020.9093390. URL https://ieeexplore.ieee.org/document/9093390.

[37] Boyan Wei, Xianfeng Ye, Chengjiang Long, Zhenjun Du, Bangyu Li, Baocai Yin, and Xin Yang. Discriminative active learning for robotic grasping in cluttered scene. *IEEE Robotics and Automation Letters*, 8(3):1858–1865, 2023. doi: 10.1109/LRA.2023.3243474. URL https://ieeexplore.ieee.org/abstract/document/10041756.

[38] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 93–102, 2019. doi: 10.1109/CVPR.2019.00018. URL https://ieeexplore.ieee.org/document/8954021.

[39] Weiping Yu, Sijie Zhu, Taojiannan Yang, and Chen Chen. Consistency-based active learning for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3950–3959, 2022. doi: 10.1109/CVPRW56347.2022.00440. URL https://ieeexplore.ieee.org/document/9857118.

[40] Hanbo Zhang, Xuguang Lan, Xinwen Zhou, Zhiqiang Tian, Yang Zhang, and Nanning Zheng. Visual manipulation relationship network for autonomous robotics. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 118–125, 2018. doi: 10.1109/HUMANOIDS.2018.8625071. URL https://ieeexplore.ieee.org/document/8625071.

[41] Hanbo Zhang, Xuguang Lan, Site Bai, Lipeng Wan, Chenjie Yang, and Nanning Zheng. A multi-task convolutional neural network for autonomous robotic grasping in object stacking scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6435–6442, 2019. doi: 10.1109/IROS40897.2019.8967977. URL https://ieeexplore.ieee.org/document/8967977.

[42] Hanbo Zhang, Xuguang Lan, Xinwen Zhou, Zhiqiang Tian, Yang Zhang, and Nanning Zheng. Visual manipulation relationship recognition in object-stacking scenes. *Pattern Recognition Letters*, 140:34–42, 2020. doi: https://doi.org/10.1016/j.patrec.2020.09.014. URL https://www.sciencedirect.com/science/article/pii/S0167865520303445.

[43] Hui Zhang, Jef Peeters, Eric Demeester, and Karel Kellens. A cnn-based grasp planning method for random picking of unknown objects with a vacuum gripper. *Journal of Intelligent & Robotic Systems*, 103:1–19, 2021. doi: 10.1007/s10846-021-01518-8. URL https://link.springer.com/article/10.1007/s10846-021-01518-8.

[44] Hui Zhang, Jef Peeters, Eric Demeester, and Karel Kellens. Deep learning reactive robotic grasping with a versatile

vacuum gripper. *IEEE Transactions on Robotics*, 39(2): 1244–1259, 2023. doi: 10.1109/TRO.2022.3226148. URL https://ieeexplore.ieee.org/abstract/document/9994578.

[45] Fedor Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019. URL https://arxiv.org/abs/1901.05954.

[46] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. URL http://arxiv.org/pdf/1801.09847v1.

[47] Guoyu Zuo, Jiayuan Tong, Hongxing Liu, Wenbai Chen, and Jianfeng Li. Graph-based visual manipulation relationship reasoning network for robotic grasping. *Frontiers in Neurorobotics*, 15, 2021. doi: 10.3389/fnbot.2021.719731. URL https://www.frontiersin.org/articles/10.3389/fnbot.2021.719731.