2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Scaling Object-centric Robotic Manipulation with Multimodal Object Identification

Chaitanya Mitash, Mostafa Hussein, Jeroen Vanbaar, Vikedo Terhuja, Kapil Katyal

***Abstract*— Robotic manipulation is a key enabler for automation in the fulfillment logistics sector. Such robotic systems require perception and manipulation capabilities to handle a wide variety of objects. Existing systems either operate on a closed set of objects or perform object-agnostic manipulation which lacks the capability for deliberate and reliable manipulation at scale. Object identification (ID) unlocks the ability for large-scale, object-centric manipulation by mapping object segments to one of the previously seen objects from a database. Nevertheless, it is often limited by the availability of reference data or coverage for objects in a database. In this work, we propose to perform object identification with multiple reference databases, including images and text references, each with a different coverage and matching challenge. We propose a training strategy that tackles the challenges of learning domain-invariant image embeddings, image-text matching and fusing predictions from different sources. We perform experiments over a recent benchmark with over 190K+ unique objects, extend the dataset with the additional reference sources and propose an evaluation strategy that simulates coverage for different reference sources. Model trained with the proposed learning pipeline shows robust performance over a range of simulation experiments.**

I. Introduction

There has been significant progress towards developing technologically advanced solutions in fulfillment logistics to meet the progressive growth of the e-commerce sector. A key enabler in this space is a robotic manipulation system that could automate repetitive operations within a warehouse such as sorting, picking and packing of orders. Towards building a scalable manipulation system, the community has transitioned from traditional object recognition pipelines [10], [49], which were limited to closed sets of objects, to object-agnostic grasp learning [27], [48], [38] and manipulation at the category-level [28]. Such approaches lose out on the benefits of object-level manipulation. An alternative is a scalable Object Identification (ID) system. As shown in Fig.1, Object ID allows a) retrieving exact attributes of an object, such as mass, shape, size etc. for manipulation planning, b) retrieving historical context from previous encounters such as successful/failed picks, damage occurrences etc. to continuously improve efficiency and reliability of the system and, c) generalizability to more deliberate tasks such as picking a specific object from a container.

Object ID is often solved via matching a query image segment with reference images of objects stored in a database. The database contains images from past encounters with the object in a similar setting. A recently released benchmark, ARMBench [29] considers the object identification challenge in the context of warehouse robotic manipulation. It contains query images from a robotic manipulation setup and reference images of objects on a tray. While the dataset contains a wide variety of objects, it is limited by the fact that reference images are often unavailable in the database. The paper acknowledges this is a common scenario and presents it as an uncertainty estimation challenge. Even if the perception system can effectively model uncertainty and make predictions when reference images are available, its utility is limited if the database lacks comprehensive object coverage.

This work addresses the above issue by performing object identification over multiple reference databases, each with a partial coverage of objects. A key insight is that objects are imaged and varying levels of detail are recorded for different purposes in the fulfillment process. Although individual reference sources may only offer partial object coverage and limited details about those objects, the combination could provide an effective representation for object identification.

Fig. 1. Cluttered container with a variety of objects in a typical warehouse setting. Object identification maps detected image segments to previously seen objects by matching with reference images. This work generalizes object identification to operate with multiple reference databases, each with partial coverage of objects. This unlocks the ability for performing more deliberate tasks such as targeted picking and leveraging stored attributes or historical context of objects for robotic manipulation at scale.

The authors are with Amazon Robotics, MA, USA. {cmitash, mosthuss, jeroenvb, terhuja, kkatyal}@amazon.com

Specifically, this paper has the following contributions:

- We present the Object Identification challenge within a novel framework that involves multiple reference sources, each with a partial object coverage. We believe this will enable object-level manipulation at scale, without assumptions of a perfect object database.
- We propose a three stage training strategy to this challenge that comprises learning domain-invariant image embedding, contrastive learning for image-text matching and learning to fuse predictions from different sources. We discover and leverage the surprising cross-modal retrieval abilities of learned embeddings even without explicit supervision. The multimodal ID solution demonstrates robustness over a large-scale dataset with a variety of reference coverage scenarios.
- We extend the existing ARMBench dataset [29] with additional reference sources and an evaluation procedure that simulates combinations of reference databases with partial object coverage to validate the robustness of Object ID algorithms.

II. Related Work

Autonomous manipulation leveraging scalable solutions has been studied extensively in prior work. In 2015, the first Amazon Picking Challenge [10] was introduced with the goal of defining real-world challenges in warehouse automation and integrating state-of-the-art algorithms in perception, motion/grasp planning and high level task planning. In the first two years of the challenge, majority of the approaches leveraged object segmentation [23] or bounding-box detection followed by pose estimation [34], [18], [49] with 3D models to compute grasps. In the following years, the trend shifted towards learning object-agnostic grasps or affordances [24], [39], [27], [30], [48] as it would better allow generalization to novel objects. In the absence of object recognition, such pipelines resort to task specification at the category-level [28] or via language [35]. However, such approaches are hard to scale in warehouse automation scenario with a wide variety of objects and tasks being specified at an instance-level. A recent benchmark, ARMBench [29] presents some of the challenges in such a setup. The current work focuses on a generalized formulation for the object identification problem that is key to unlocking object-centric manipulation at scale.

The Object ID task is similar to that of image retrieval, where given a query image, the most similar image is retrieved from an image database. This task has been studied in the context of landmark recognition [45], fashion [25], product retrieval in e-commerce [8], [50], and person re-ID [47]. Previous approaches on this task consider aggregating pre-defined local features [37], [31], [21] and computing similarity metrics over features derived from large-scale image classification training [2], [41], [15]. More recent approaches consider metric learning via siamese [9], [20], [36] or triplet networks [43], [19]. These approaches use matching and non-matching pairs of images to train features for image retrieval, with additional tricks like deep local feature aggregation [44], [40] and using margin-based loss [11]. The problem considered in our work pose additional challenges such as high precision requirement, large occlusions and presence of multiple domains with different notions of similarity which demand learning a generalizable representation.

Several recent work [3], [6], [7], [16], [17], [4] have shown the effectiveness of self-supervision in learning image representations that provide zero-shot transfer properties to different tasks and domains. Such approaches pair different augmentations of the same image to train image embeddings. Of these, the DINO [4] approach uses a knowledge distillation paradigm to train a vision transformer [14] with image augmentations that include multiple local crops of the image. DINO embeddings have shown interesting properties [1] and improvements in a variety of tasks including nearest neighbor retrieval for landmarks as well as video segmentation. Given the ability of this framework to generalize across domains and capture local-to-global correspondences, we adapt it as a visual backbone for object identification.

With the emergence of VL models, multimodal retrieval has gained popularity. Recent works often use Contrastive Language–Image Pre-training (CLIP) [32], a transformer-based architecture for image and language encoding, which is trained on image-text pairs in a contrastive setting. Some work specifically address the challenges of multimodal retrieval under ambiguities. In [46], the authors propose a so-called Retrieval Augmented Module, to capture additional images with text descriptions based on the query image to enrich the input representation. In our case we typically have access to only a single text description for an object. The authors of [26] recognize the ambiguity of certain words between the general and product domains, and aim to capture domain semantics via entities. Their work is currently limited to the fashion domain. Text can be leveraged to retrieve and localize instances between different images of the product [22]. This does not address large variation in appearance, for example between images of the actual product and its packaged version. Leveraging a plurality of modalities can be beneficial for retrieval in the context of e-commerce [13], however we currently only have access to images and text. Given our need to handle a large variety of objects, with potentially large appearance difference, and partial coverage of some reference data, we will show that we can use a DINO model trained for domain-invariant embeddings as the image encoder for CLIP, and finetune using an extended version of the ARMBench dataset [29] to align images and text.

III. Problem Setup

This work considers the problem of Object Identification in the context of a common pick-and-place robotic manipulation setup (Fig. 1) with the following inputs:

- A query image depicting an object segment, typically output of instance segmentation over a cluttered scene.
- A candidate set that comprises a list of possible objects. This is often the list of all objects in a cluttered container, derived via inventory tracking systems.
- A gallery that contains a set of reference data for candidate objects.

Fig. 2. Examples of the four different reference sources used for object identification. The Tray and Bin images are acquired during fulfilment operations and are representative of packaged state of objects while Catalog images and text are descriptions of the object typically used in an online catalog or webpage.

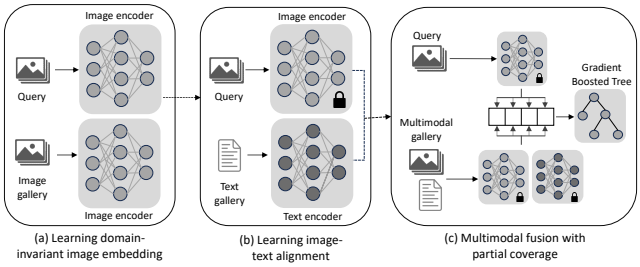

Fig. 3. We propose a solution for multimodal identification under partial coverage of reference sources consisting of three stages. In the first stage we train an image encoder in a combination of self-supervision and supervision over multiple reference sources to obtain a domain-invariant embedding that is suitable for retrieval. The second stage aims to align a text embedding with the image embedding from the first stage, by finetune training over image-text pairs in a contrastive setting. Finally, a decision tree is trained on embedding distances in a third stage, in order to achieve highest multimodal ID performance for galleries of multiple reference sources with partial coverage.

Object identification is then performed via matching the query image with gallery data to output the most-likely object from the candidate set and an accompanying confidence value.

For the case where we have unimodal, single reference source [29], this task is similar to person re-ID [47]. Reference images of the object may have different viewpoints and somewhat different appearance. However, for our problem domain, coverage for a single reference source at inference time can be well below 100%, in some cases as low as 70%. To counter this issue of missing data for some objects, this work introduces the idea of leveraging multimodal data from multiple reference sources for the gallery. Specifically, it considers four different sources, namely, Tray images, Bin images, and a Catalog image and title.

- Tray images – Images of the packaged object, from various viewpoints, acquired at some point prior to arrival at the pick-and-place robotic manipulation station.
- Bin images – Images of the packaged object acquired while in storage during the fulfillment logistics process.
- Catalog image – A representative image of the object, typically used in an online catalog.
- Catalog title – Text description for an object to accompany the catalog image.

Some examples are shown in Fig. 2. Images of objects from multiple gallery sources may have substantially different appearance, including lower resolution and partial occlusion by translucent bands, as in the case with Bin images. Additionally, the Catalog images and title of objects may have very sparse and low semantic association with the images of objects captured during processing in a fulfilment center.

In this paper, we aim to address the challenges of partial coverage by exploiting multimodal data from multiple reference sources for object identification. We propose a solution consisting of three stages (see Fig. 3): learning domain-invariant image embedding via global-to-local correspondence training with multiple reference sources, learning image-text alignment in a joint embedding space via contrastive learning, and finally a fusion approach for matching based on distances computed using different embedding spaces. We explain challenges of each stage, and motivation behind our multi-stage approach, in subsequent sections.

In order to train and test the models of the three aforementioned stages, we have extended the ARMBench dataset [29] with data from the alternate gallery databases as mentioned above. The dataset contains ∼235K pick activities, consisting of ∼190K unique items. On average, an item has 4x Tray images, 5x Bin images and a single Catalog image with description of item. The main evaluation metric we use in the paper is Precision@X% ID rate, where ID rate corresponds to all cases where an ID prediction is made. In order to compare ID performance for the different reference sources, we report the performance for individual sources as Precision@100% ID rate, or top-1 retrieval rate. When we discuss the fusion approach, we also evaluate for combined sources. In this case we sample candidate sets of possible objects and simulate different coverage scenarios, both according to estimates from real-world applications, as well as coverage scenarios for ablation studies. Performance in this case is shown as Precision vs ID rate plots.

The following sections discuss the key challenges of a) matching query images with reference images from different domains, b) matching query images with text description of objects and c) fusing the outcome of matching with different reference sources in the presence of partial coverage.

IV. Domain-Invariant Image Embeddings

Learning embeddings that allow comparing images by separating similar and dissimilar pairs has emerged as a popular strategy in open set recognition and image retrieval setting. Nevertheless, such an approach is further challenged in this problem setup due to the presence of multiple image domains, each with its unique criteria for similarity to the query image.

To solve this challenge, we adapt the DINO framework [4] that combines knowledge distillation with self-supervised learning to train vision transformers. As shown in Fig. 4(b), DINO is typically trained by passing two different transformations of the same image via a student

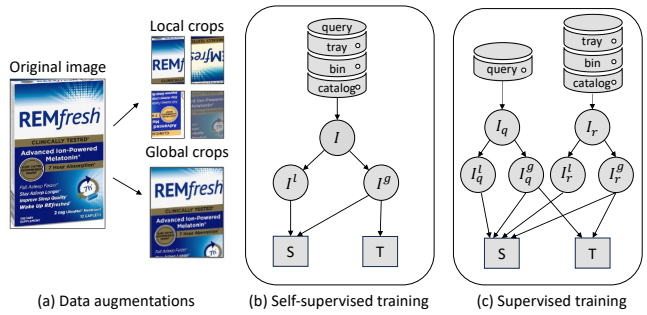

Fig. 4. (a) Different augmentations used for training: local (20-40%) and global crops (50-90%), affine transformation, color jittering. (b) Self-supervised model trained via different augmentations of the same image. I^l corresponds to a local crop and I^g is a global crop of the image. S and T are student and teacher networks respectively. (c) Supervised learning where corresponding images are sampled from query (I_q) and reference images (I_r) and subjected to augmentations.

and a teacher network and then matching the similarity between the outputs via a cross-entropy loss. Fig. 4(a) shows examples of these transformations, including the global (50-90%) and local crops (20-40%) of the original image size. Training with these crops encourages learning local-to-global correspondences. We first evaluate the effectiveness of this self-supervised training policy for object ID in our scenario. We compare two models (first two rows of Table I), where the first model is trained on the ImageNet dataset [33] and we trained the second model on the ARMBench training dataset. For this training, we used query as well as all image reference sources to train a ViT-S/16 for 300 epochs. We observe that self-supervised training on the ARMBench dataset improves top-1 retrieval rate across all image domains. While the retrieval rate with tray images goes up to 85.2%, it is much lower, 42.6% and 36.6% respectively for bin and catalog images, indicating a larger domain-gap.

To address this domain-gap, we leverage the supervision from the ARMBench training set that provides correspondences between query and reference images for over 150K unique objects. For each iteration, we sample an object and a random corresponding image from one of the reference sources. Global crops of these images are passed to the teacher network, while both global and local crops are passed to the student network. The network is trained this time with cross-domain association. We additionally performed experiments with sampling only global crops of the images based on the intuition that local crops of images from different domains and viewpoints might have no shared context. Both of these training modes were initialized using the weights from the self-supervised model trained on ARMBench dataset. Table I, rows 3 and 4 show the outcome of these experiments. The improvement from the supervised learning step is significant, and contrary to our intuition, we observe that using local crops further improves the retrieval rate, especially for bin and catalog images. This could be attributed to the fact that both of these reference sources only have a small part of the image that contains the relevant information to match with query images and thus can benefit from learning local-to-global correspondences.

TABLE I

TOP-1 RETRIEVAL RATE FOR DIFFERENT REFERENCE SOURCES

training	tray	bin	catalog
DINO-Self-Supervised (ImageNet [33])	70.8	23.8	27.6
DINO-Self-Supervised (ARMBench [29])	85.2	42.6	36.6
DINO-Supervised (global crops)	96.8	90.5	60.7
DINO-Supervised (global+local crops)	97.7	93.7	64.5
DINO-Supervised-incremental	97.8	94.0	68.1

Finally, inspired by Curriculum learning [52] where a model is trained on a progressively challenging sequence of tasks or examples, we train our model in an incremental fashion i.e. tray first, followed by tray and bin and followed by tray, bin and catalog. Using this technique achieves the best result so far, particularly for catalog images as shown in Table I, Row 5.

We observe that, even with our best model, the ID performance using the catalog gallery source (68.1%) is significantly lower compared to bin or tray (94.0%, and 97.8%) images. We believe this is because catalog images are designed to describe the product and only a small part of the image is relevant to matching with the query image. In addition to image data from reference databases, we also have access to the object's textual descriptions via catalog titles. To address the lower performance on catalog images, we investigate matching query images with catalog titles as it may contain the relevant information in a lower dimensional space.

V. LEARNING IMAGE-TEXT EMBEDDINGS

Vision-Language (VL) models have gained popularity recently for tasks such as image and text retrieval, visual question and answering, and commanding robots with natural language [51]. In our work, we chose CLIP [32] as the VL model. CLIP requires image-text pairs as training data, and is then subsequently trained in a self-supervised contrastive setting. The CLIP architecture is a *two-tower model*, consisting of an image encoder and a text encoder. The text encoder is a BERT-like transformer [12]. Although the image encoder can be either a residual network or ViT, in this paper we only consider the latter based on both performance and ability to compare with and exploit the DINO ViT. CLIP aims to preserve mutual information between the image and text embeddings, by using a noise-contrastive estimation loss (InfoNCE [42]). The goal for training CLIP is to minimize the distance in the embedding space between the paired image and text embeddings, while simultaneously maximizing the distance between non-paired images and text embeddings, depicted in Fig. 5.

As mentioned in [32], pre-trained CLIP models do not perform well on out-of-distribution data. Since our query images of segmented objects fall into that category, we use the ARMBench dataset to finetune the pre-trained CLIP models. Table II shows that finetuning gives around a 35-40% boost in performance. We note that although the ViT-L/14@336 variant provides the highest performance, it also has about $8\times$ the number of trainable parameters compared

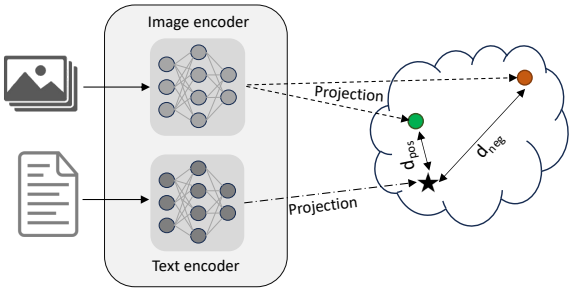

Fig. 5. CLIP takes paired images and text, and trains an image and text encoder in a contrastive setting. Imposing a contrastive loss aims to minimize the distance in the embedding space between the paired image and text embeddings (d_{pos}), while simultaneously maximizing the distance between non-paired images and text embeddings (d_{neg}).

TABLE II

TOP-1 RETRIEVAL RATE FOR CLIP VARIANTS, FOR BOTH PRE-TRAINED AND FINETUNED MODELS.

Embedding Models	Pre-trained	Finetuned
ViT-B/16	42.7	82.1
ViT-B/32	34	75.1
ViT-L/14	50	85.1
ViT-L/14@336	52	86.5

to the ViT-B/16 variant, which is an important deployment consideration in terms of computational resources.

As a side note, Catalog titles often contain additional information about the object, such as dimensional information, or information related to function rather than appearance. We postulate that this additional title information may hinder the embedding learning, since there is no relevant image information directly associated with this. We truncate the titles on punctuation marks such as commas, semi-colons, or pipe symbols. Our experiments show that title truncation provides around a 2% performance improvement, and all numbers reported in the paper use this truncation.

TABLE III

TOP-1 RETRIEVAL RATE FOR VISUAL AND TEXT MATCHING

training	tray	bin	catalog	text
CLIP-pretrained	60.1	20.7	34.6	42.7
CLIP-finetuned	90.8	47.1	66.3	82.1
DINO-image-text-finetuned	95.3	87.3	58.3	78.9
DINO-text-finetuned	97.8	94.0	68.1	80.6

Table II shows only results for Catalog title however, we want to evaluate the CLIP model across the multiple references in order to compare with the results in Table I. We can use the image encoder of the CLIP model directly to perform the same visual ID as in the previous section. Table III, rows 1 and 2, shows visual ID performance for the *CLIP-pretrained* and *CLIP-finetuned* ViT-B/16 variant on tray, bin and catalog gallery reference sources. Surprisingly, even though the finetune training only learns to align query images and catalog titles, the visual ID performance of the finetuned CLIP model on tray and catalog images is quite

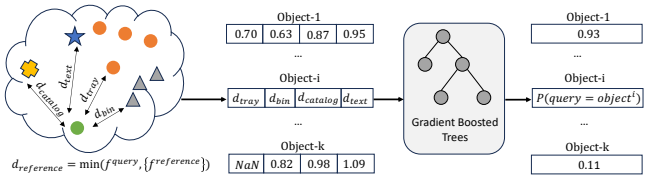

Fig. 6. For every object in the candidate set, a minimum distance is computed between the query image embedding and embeddings from each reference source. This 4d vector of distances for an object is then input to a gradient boosted tree that returns the probability of the query image corresponding to that object. Above, *NaN* represents missing data due to partial coverage.

good when compared with DINO. The performance on bin images on the other hand, is much lower compared with DINO, and we attribute this to the relatively large difference in appearance between query and bin images.

Since the *Supervised-incremental* DINO model from Sec. IV already shows good performance on the different reference sources, and the encoder for DINO is a ViT-S/16, we propose to use this as the pre-trained image encoder in the CLIP framework. By finetuning CLIP with this pre-trained image encoder, we hope to achieve good performance for both images and text. We use two finetuning regimes: one where we finetune both the image and text encoder (labeled as *DINO-image-text-finetuned*) and another where we freeze the weights of the DINO pretrained ViT-S/16 and finetune only the text encoder (labeled as *DINO-text-finetuned*). For *DINO-image-text-finetuned*, in addition to query-text, we add contrastive loss terms for tray-text, bin-text, catalog-text, query-tray, query-bin and query-catalog image. For *DINO-text-finetuned* we omit the projection layer for the CLIP image encoder, but project the text encodings from the CLIP text encoder to match the lower resolution of the ViT-S/16 embedding space. We note that the text projection is implemented as a linear layer with learnable parameters.

When comparing rows 3 and 4 in Table III, we see that *DINO-text-finetuned* gives the best performance, and retains the visual ID performance of the best model in row 5 of Table I, with only slight performance difference for ID using text compared to *CLIP-finetuned* (Table III, row 2). It is evident that the *Supervised-incremental* pre-training of the image encoder for visual ID, has captured the necessary information required for subsequent image-text alignment finetuning within the CLIP framework. An added benefit in terms of deployment is the fact that the number of parameters of the ViT-S/16 are about $0.5\times$ that of a ViT-B/16.

VI. MULTIMODAL IDENTIFICATION WITH PARTIAL GALLERY COVERAGE

The experiments in the previous two sections focused on evaluating ID retrieval rate with individual reference sources. However, in the real-world scenario, inference needs to be made over the entire gallery, where each object will have images from one or more reference sources. This is challenging because the distances of query images to different reference sources are often not calibrated. Fig.7 (left) shows

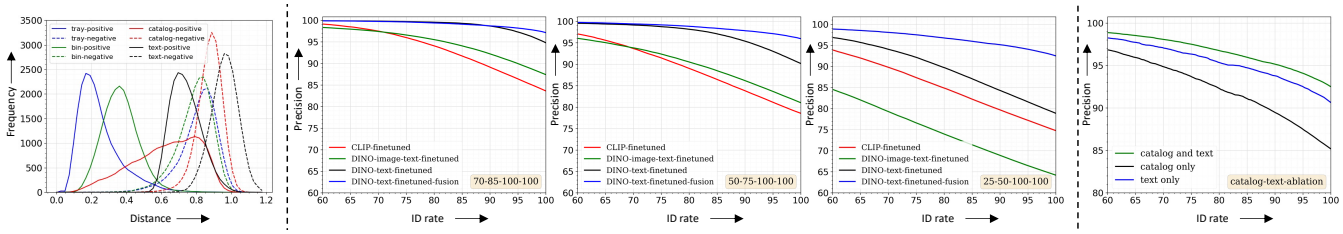

Fig. 7. (left) frequency plot for query-reference distance shows the need for distance calibration across modalities (middle) plots show the precision vs ID rate plot in different coverage scenarios. The coverage is indicated as tray-bin-catalog-text (right) shows the impact in precision by eliminating text or catalog images with tray, bin coverage set to 25 and 50 respectively.

Fig. 8. Failure cases: (Top) the query image, (Bottom) highest confidence match. In all these cases, the correct item only had catalog image as reference source which was not representative of the appearance.

the frequency of distances of different reference sources with query images for positive and negative object samples that highlights this issue.

To address the calibration issue, we train a fusion model (Fig. 6) that takes as input minimum distances between the query image and each of the reference sources for a specific object and predicts the probability of the query image being that object. We use the XGBOOST framework [5] to train a gradient boosted tree for this fusion task. The training data comprises 25K query images with corresponding ground-truth reference images. The data is augmented with large number of variations in reference coverage and negative samples to generate approximately 500K data points for training. Distance value corresponding to cases of no coverage is set to NaN both during training and inference.

Evaluation is performed over 100K test cases corresponding to 25K query images. For each testcase, a candidate set is sampled which contains 10-30 objects sampled from a collection of approximately 30K unique objects. This set is chosen so that each object has 100% reference coverage for all sources. Then based on a pre-specified config, reference coverage is simulated for this candidate set. E.g., a reference coverage of 70-85-100-100 will simulate a 70% coverage for tray images, 85% for bin, and 100% for catalog and text.

Fig. 7 (middle) plots the precision of models as a function of the ID rate (total % of cases where a prediction is made) for three different reference coverage scenarios. For all models except the fusion model, prediction is based on the closest reference data to query image irrespective of the source. The plots indicate the efficacy of the clustering achieved via dino-text-finetuned model compared to alternate methods. Nevertheless, even for that model the precision drops significantly for lower coverage scenarios. The fusion approach applied over this model provides additional robustness. It achieves 97.1%, 95.9% and 92.4% ID retrieval rates for the three reference coverage scenarios. Finally, in Fig. 7 (right), we evaluate the contribution of catalog and text sources to the fusion model. The plot indicates that the two sources indeed complement each other.

Fig. 8 shows some examples of high-confidence failure cases with the fusion model that can result in false positives. A common failure we notice corresponds to white bags and brown boxes where the actual object is packaged inside. In both of these cases (Fig. 8), the only available reference source for the ground-truth object was catalog image and title while the predicted object additionally had access to tray images. The tray image for the predicted object happened to indicate a very similar packaging that could have resulted in the high-confidence prediction.

VII. Discussion and Future Work

This paper tackles the challenge of missing reference data for object identification in a large scale robotic manipulation setup. The solution demonstrates that multiple, partial reference databases, across different image domains and modalities can be simultaneously used for this task. We discover that a local-to-global correspondence training with self-distillation can be used in a supervised setting with images from very different domains, and that it can even capture text features that allow object retrieval. Based on this learning, we propose a solution that learns domain-invariant image embeddings and then projects text features into this space via contrastive learning. While the projection already enables multimodal retrieval via computing a nearest neighbor in the combined space, we find that calibrating the distances across modalities using a fusion step makes the retrieval more robust to scenarios with varying reference coverage. All the experiments are performed on the ARMBench dataset that contains over 190K+ unique objects. We extend this dataset with reference images from the additional sources used in this paper.

Formulating the object identification problem with multiple reference sources opens the avenue for a wide range of future research. This includes: improving the retrieval rates for matching with individual reference sources such as the bin and catalog images, finding ways to simultaneously train image and text encoders while keeping the advantages of local-to-global image training, and, incorporating the fusion of modalities within the feature learning pipeline.

REFERENCES

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2022.

[2] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *European conference on computer vision*, pages 584–599. Springer, 2014.

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *ArXiv*, abs/2006.09882, 2020.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[5] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.

[7] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2020.

[8] Lele Cheng, Xiangzeng Zhou, Liming Zhao, Dangwei Li, Hong Shang, Yun Zheng, Pan Pan, and Yinghui Xu. Weakly supervised learning with side information for noisy labeled images. *ArXiv*, abs/2008.11586, 2020.

[9] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:539–546 vol. 1, 2005.

[10] Nikolaus Correll, Kostas E Bekris, Dmitry Berenson, Oliver Brock, Albert Causo, Kris Hauser, Kei Okada, Alberto Rodriguez, Joseph M Romano, and Peter R Wurman. Analysis and observations from the first amazon picking challenge. *IEEE Transactions on Automation Science and Engineering*, 15(1):172–188, 2016.

[11] Jiankang Deng, J. Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2018.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[13] Xiao Dong, Xunlin Zhan, Yangxin Wu, Yunchao Wei, Michael C. Kampffmeyer, Xiaoyong Wei, Minlong Lu, Yaowei Wang, and Xiaodan Liang. M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21220–21230, 2022.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[15] Noa Garcia and George Vogiatzis. Learning non-metric visual similarity for image retrieval. *Image and Vision Computing*, 82:18–25, 2019.

[16] Jean-Bastien Grill, Florian Strub, Florent Altch'e, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *ArXiv*, abs/2006.07733, 2020.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2019.

[18] Carlos Hernandez, Mukunda Bharatheesha, Wilson Kien Ho Ko, Hans Gaiser, Jethro Tan, Kanter van Deurzen, Maarten de Vries, Bas Van Mil, Jeff van Egmond, Ruben Burger, Mihai Morariu, Jihong Ju, Xander Gerrmann, Ronald M. Ensing, Jan van Frankenhuyzen, and Martijn Wisse. Team delft's robot winner of the amazon picking challenge 2016. In *Robot Soccer World Cup*, 2016.

[19] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, 2014.

[20] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1882, 2014.

[21] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010.

[22] Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Learning instance-level representation for large-scale multi-modal pretraining in e-commerce. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11060–11069, 2023.

[23] Rico Jonschkowski, Clemens Eppner, Sebastian Höfer, Roberto Martin Martin, and Oliver Brock. Probabilistic multi-class segmentation for the amazon picking challenge. *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–7, 2016.

[24] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34:705 – 724, 2013.

[25] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016.

[26] Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18051–18061, June 2022.

[27] Jeffrey Mahler, Matthew Matl, Vishal Satish, Michael Danielczuk, Bill DeRose, Stephen McKinley, and Ken Goldberg. Learning ambidextrous robot grasping policies. *Science Robotics*, 4(26):eaau4984, 2019.

[28] Lucas Manuelli, Wei Gao, Peter Florence, and Russ Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In *The International Symposium of Robotics Research*, pages 132–157. Springer, 2019.

[29] Chaitanya Mitash, Fan Wang, Shiyang Lu, Vikedo Terhuja, Tyler Garaas, Felipe Polido, and Manikantan Nambi. Armbench: An object-centric benchmark dataset for robotic manipulation. *arXiv preprint arXiv:2303.16382*, 2023.

[30] Douglas Morrison, Adam W. Tow, M. McTaggart, R. Smith, N. Kelly-Boxall, S. Wade-McCue, J. Erskine, R. Grinover, A. Gurman, T. Hunn, D. Lee, Anton Milan, Trung T. Pham, G. Rallos, Andrew Razjigaev, Thomas Rowntree, B. V. Kumar, Zheyu Zhuang, Christopher F. Lehnert, Ian D. Reid, Peter Corke, and J. Leitner. Cartman: The low-cost cartesian manipulator that won the amazon robotics challenge. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7757–7764, 2017.

[31] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[34] Max Schwarz, Anton Milan, Arul Selvam Periyasamy, and Sven Behnke. Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter. *The International Journal of Robotics Research*, 37:437 – 451, 2018.

[35] Bokui (William) Shen, Ge Yang, Alan Yu, Jan Rang Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. *ArXiv*, abs/2308.07931, 2023.

[36] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, P. Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 118–126, 2015.

[37] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *null*, page 1470. IEEE, 2003.

[38] Andreas Ten Pas, Marcus Gualtieri, Kate Saenko, and Robert Platt. Grasp pose detection in point clouds. *The International Journal of Robotics Research*, 36(13-14):1455–1473, 2017.

[39] Andreas ten Pas, Marcus Gualtieri, Kate Saenko, and Robert W. Platt. Grasp pose detection in point clouds. *The International Journal of Robotics Research*, 36:1455 – 1473, 2017.

[40] Giorgos Tolias, Tomás Jenícek, and Ondvrej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *European Conference on Computer Vision*, 2020.

[41] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015.

[42] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

[43] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.

[44] Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Learning super-features for image retrieval. *ArXiv*, abs/2201.13182, 2022.

[45] Tobias Weyand, Andre F. de Araújo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2572–2581, 2020.

[46] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19265–19274, June 2023.

[47] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *CoRR*, abs/2001.04193, 2020.

[48] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *The International Journal of Robotics Research*, 41(7):690–705, 2022.

[49] Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker, Alberto Rodriguez, and Jianxiong Xiao. Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 1386–1383. IEEE, 2017.

[50] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11762–11771, 2021.

[51] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*, 2023.

[52] Jiwen Zhang, Jianqing Fan, Jiajie Peng, et al. Curriculum learning for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34:13328–13339, 2021.