Object Recognition with Continual Open Set Domain Adaptation for Home Robot

Ikki Kishida[1], Hong Chen[1], Masaki Baba[1], Jiren Jin*[1], Ayako Amma[2], and Hideki Nakayama[1]

[1] *Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan*
[2] *Toyota Research Institute - Advanced Development, Tokyo, Japan*
{kishida,chen,baba,jin,nakayama}@nlab.ci.i.u-tokyo.ac.jp[1], ayako.amma@tri-ad.global[2]

Abstract

Object recognition ability is indispensable for robots to act like humans in a home environment. For example, when considering an object searching task, humans can recognize a naturally arranged object previously held in their hands while ignoring never observed objects. Even in such a simple task, we need to deal with three complex problems: domain adaptation, open-set recognition, and continual learning. However, most existing datasets are simplified to focus on one problem and do not measure the object recognition ability for home robots when multiple problems are simultaneously present. In this paper, we propose the COSDA-HR (Continual Open Set Domain Adaptation for Home Robot) dataset that requires dealing with the above three problems simultaneously. The COSDA-HR dataset focuses particularly on the scenario in which naturally arranged objects in a room are recognized by training with handheld objects towards the goal of creating a user-friendly teaching system for home robots. We provide various baselines to address the problems in the COSDA-HR dataset by combining state-of-the-art methods from each research area and analyze the limitations of such simple combinations. We consider that it is necessary to study the methods of handling multiple problems simultaneously instead of solving each problem to realize practical object recognition systems for home robots.

1. Introduction

Home-assistant robots are one of the applications where visual recognition plays a vital role. For example, for object retrieval or tidying-up tasks in a home environment, a robot is required to recognize multiple objects in cluttered environments. With the rapid progress of visual recognition techniques, the development of such cognitive robots has been gaining considerable attention and expectation. However, to put a robot in a home environment with a realistic application scenario, how to train the systems remains a big challenge. Because the environment and types of objects considerably vary from home to home, it is very difficult to pre-train a recognition system that is universally applicable to all homes. Therefore, we need an efficient framework to train the recognition ability of robots in each environment without much burden for end-users. We consider that displaying objects in front of robots is one of the user-friendly scenarios of teaching objects for a visual recognition system in home robots.

To attain this scenario more realistically, we claim that the following three aspects are important and considered simultaneously. **(i) Domain adaptation**: it is difficult and not realistic for end-users to collect sufficient training data (images and annotations) in a real target environment (e.g., a desk or a shelf), which is often complex and messy. A more controllable and feasible way of teaching a robot would be to show an object one by one in front of its eye camera. However, this means that a huge domain shift inevitably occurs between the source domain (handheld object) and the target domain. **(ii) Open-set recognition**: in a real environment, there should be a number of unknown objects not presented in the training phase. The recognition system must be able to distinguish such unknown objects from known ones. **(iii) Continual learning**: as novel objects continually appear in our daily life, the recognition system should be able to update the knowledge timely to handle them. A naive strategy is to keep all previous training data and retrain the system from scratch every time new data are added. However, this approach is extremely expensive in terms of computation and memory costs, and not feasible in practical user-side systems. Therefore, we need an efficient frame-

*now at Google.

Dataset	Handheld	OS	DA	CL	Robot eye-level	Depth	# Objects (# Categories)	# Images
ROD [17]			✓			✓	300 (51)	250,000
ARID [23]			✓		✓	✓	153 (51)	28,316
OpenLORIS-object [37]				✓	✓	✓	69 (19)	430,560
CORe50 [24]	✓			✓		✓	50 (10)	164,866
iCWT [30]	✓		✓		✓	✓	200 (20)	414,483
COSDA-HR (ours)	✓	✓	✓	✓	✓	✓	160 (16)	89,339

Table 1: Properties of datasets to evaluate object recognition ability of robots. Our proposed COSDA-HR dataset is the first attempt to measure the object recognition ability of a home robot when recognizing naturally arranged objects in a room by training with handheld ones in a realistic situation where in the three problems of domain adaptation (DA), open-set recognition (OS), and continual learning (CL) are simultaneously present.

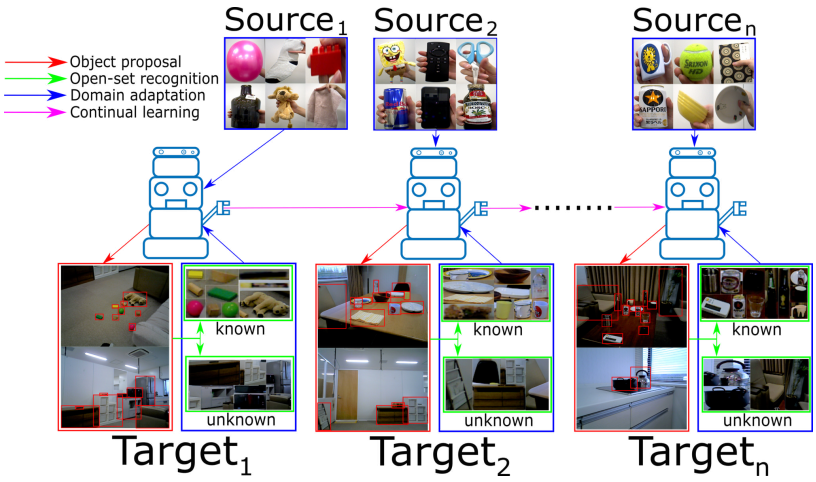

Figure 1: Overview of the task pipeline in the COSDA-HR dataset. The system is given some labeled handheld images as the training data of the source domain while it also explores the environment and detects object-like regions from unlabeled target images. Using the source and target data, the system performs domain adaptation and recognition under the open-set situation. This procedure is incrementally repeated using n different subsets of classes.

work to incrementally train the system by only using the novel data without forgetting the knowledge obtained in the past.

Although there have been many sophisticated datasets for visual recognition on robots [17, 23, 37, 24, 30], to the best of our knowledge, none of them have covered the above three problems simultaneously to recognize naturally arranged objects in a room by training with handheld ones for the goal of creating a user-friendly teaching system for home robots (Table 1). As we have discussed, we consider that it is essential to consider those problems simultaneously and this is also important from theoretical viewpoints because the assumption of each specific problem is sometimes contradictory to others.

On the basis of this motivation, in this paper, we propose the COSDA-HR (Continual Open Set Domain Adaptation for Home Robot) dataset[1] as the benchmark of object recognition ability of robots in a home environment with a continual training and testing scenario (Figure 1). The assumed scenario is that we present a robot with novel handheld objects every day for training, and the robot adapts the visual knowledge from the handheld images to the home environment by using the images explored in the home on that day where many unknown objects also exist. We build the baseline models by combining state-of-the-art methods in each of the three problems and conduct several experiments on the COSDA-HR dataset. We demonstrate that a simple combination can only achieve 17% mean class accuracy in known classes when the recall for the unknown class is larger than 90%. This result implies that it is necessary to study algorithms to deal with multiple problems simultaneously instead of solving each problem separately. In summary, our contributions are as follows:

- We newly propose the COSDA-HR dataset. It is the first attempt to measure the object recognition ability of a home robot when recognizing naturally arranged

[1]The COSDA-HR dataset and the codes for the baseline models will be publicly available after the publication of this paper.

objects in a room by training with handheld ones in a realistic situation wherein the three problems of domain adaptation, open-set recognition, and continual learning are simultaneously present.

- We build the baseline models by thoroughly implementing the state-of-the-art methods for each research field of the three problems and analyze the limitations of their simple combinations.

2. Related Work

2.1. Domain Adaptation

Domain shift, which is a discrepancy between the source domain and the target domain, is known to significantly degrade the performance of learning models. Domain adaptation is the technique of alleviating the problem of the domain shift, which was firstly applied to visual object recognition in [34], and then various approaches have been proposed. Instance-based methods [12] use data importance weighting or class importance weighting to find samples in the source domain that are more related to the target domain. Feature-based methods either convert data from one domain to another or try to learn a domain-invariant representation to reduce the domain discrepancy. Recently, methods based on deep neural networks have achieved significant progress in the above approaches. For example, various discrepancy-based approaches [25, 40, 52, 38] are applied to deep features. Moreover, with the success of Generative Adversarial Networks (GAN) [9], domain adversarial training [7, 49, 43, 4, 26, 31] has become one of the dominant approaches in the field.

Nevertheless, solely aligning the data distributions does not guarantee the performance on the target domain owing to the possibility that the conditional distributions may not be aligned between the features in each domain. Therefore, several approaches [46, 44, 53, 6, 35, 18, 13, 54] are proposed to alleviate the insufficiency of global domain alignment.

2.2. Open Set Problems

Open-set recognition: The goal of open-set recognition is to detect and reject unknown-class samples while maintaining the recognition performance for known classes. In recent years, deep open-set classifiers have become the standard approach for open-set recognition. The OpenMax method [3] modifies the prediction based on the Weibull distribution trained with respect to the distance from the mean vector of each class. As the extensions of OpenMax, G-OpenMax [8] and CROSR [47] are proposed. Both methods take advantage of unsupervised learning.

Open-set domain adaptation: Open-Set Domain Adaptation (OSDA) is the mixed problem of open-set recognition and domain adaptation, and it is firstly proposed in [29]. Unlike the standard domain adaptation, in OSDA, there are some unknown classes in the target domain that do not appear in the source domain. OSDA is considered to be a more challenging but practical setting since it is difficult to guarantee that none of an unknown class appears in the target domain. To realize OSDA, [29] assumed that the source data contain the unknown class, but it is not always a feasible approach since it is difficult to define the contents of an unknown class and collect sufficient data in advance. Therefore, [36] extended the method so that it does not require an unknown class data in the source domain by introducing adversarial training. Since then, many OSDA methods without an unknown class in the source domain are proposed [48, 2, 22, 16, 28].

2.3. Continual Learning

Continual learning, also termed as lifelong learning, has attracted the attention of many researchers in the last few years. Unlike the standard offline learning, training data of novel classes are incrementally given in continual learning. The main difficulty is that we have to realize an incremental update of the knowledge with as less access to old data as possible, which has the risk of *catastrophic forgetting*, that is the model forgets what it previously learned. Various approaches have been proposed to deal with catastrophic forgetting which can be roughly separated into two categories: (1) without memory and (2) with memory.

(1) Without memory: This is the most strict setting of continual learning where the model is only allowed to access the current data and can not retain the old data. To retain the knowledge obtained from the old data, LwF [19] applies the distillation approach. EWC [14] takes an approach to force a constraint such that the parameters useful for the old data are not largely changed at the next update. DGR [39] uses Generative Adversarial Models [9] to produce synthetic data of the previous classes. Generally, the without-memory setting is quite challenging, and it is difficult to achieve strong performance comparable to offline training.

(2) With memory: This approach relaxes the condition by allowing the model to have a small fixed-size memory to store the past data [32, 27, 5, 41, 1, 50]. Some of the old data can be saved in the memory which is replayed during the incremental training at each step.

2.4. Benchmark Dataset for Object Recognition in Robotics

We review several benchmark datasets for robotic vision tasks that are related to our work and summarize the properties of these datasets in Table 1.

RGB-D Object Dataset (ROD) [17] consists of 300 objects in 51 categories of common household objects. ROD not only contains RGB-D multiview images but also in-

cludes eight video sequences of common indoor environments. The images of 300 objects are collected using a turntable, which are comprehensive in terms of object appearance. However, in practical scenarios, such a special instrument and labor might not be practical for end-users. In the COSDA-HR dataset, we rotate an object by the hand, thus it does not require such a special instrument.

Autonomous Robot Indoor Dataset (ARID) [23] consists of 153 objects in 51 categories. ARID has images captured at the first-person view of a robot working in a home environment under different lighting conditions. It also provides web-domain images to investigate the transferability of the knowledge from the web domain to the real-world domain.

OpenLORIS-object Dataset [37] is a benchmark dataset for continual learning. It is mainly focused on continual learning and considers many environmental factors, such as illumination, occlusion, clutter, object size, and camera pose in various indoor environments including home, office, coffee shop, and mole. Although the considered environmental conditions are more diverse than our dataset, OpenLORIS is not targeted to open-set and domain shift problems.

CORe50 Dataset [24] and **iCubWorld Transformations Dataset** [30] consist of many categories of handheld images such as source images in our COSDA-HR dataset. In their setting, [30] focused on object recognition in a wide variety of visual transformations, and [24] focused on object recognition in continual learning.

Why COSDA-HR dataset? It is important to provide a less burdensome way of the teaching system for end-users, and we consider that displaying handheld objects to a home robot is one of the user-friendly scenarios. When considering such a scenario in a practical situation, we consider that it is necessary to deal with the three problems of continual learning, open-set recognition, and domain adaptation simultaneously. To the best of our knowledge, all existing datasets are not suitable for measuring the recognition ability in such a scenario. ROD, ARID, and OpenLORIS dataset do not provide handheld images, thus, it is difficult to extend them for our scenario. The CORe50 dataset has the handheld objects and naturally arranged objects as test images, however, it is difficult to apply domain adaptation methods since they do not provide sufficient images in their test domains. In our scenario, domain adaptation is essential since a home environment is considerably different from home to home, and it is very difficult to pre-train a recognition system that works in all environments. The iCWT dataset also has the handheld images, however, they do not provide the naturally arranged paired objects for testing, thus iCWT could not be extended for our purpose. For the above reasons, we built the COSDA-HR dataset from scratch instead of extending the existing datasets.

3. COSDA-HR: Continual Open Set Domain Adaptation for Home Robot

3.1. Collecting Images and Annotations

All the images in the COSDA-HR dataset are captured by the Xtion RGB-D sensor mounted at the eye-level of the Toyota Human Support Robot (HSR) [42] whose appearance and specifications are shown in the supplementary material. In addition to RGB images, depth information is also available in our dataset, although we do not use it in this paper. There are 16 super categories and each category has 10 different instances, thus there are 160 classes in total. The 16 super categories are selected from those commonly appearing in a home environment, which include *ball, book, bowl, toy block, can, cup, dish, glass bottle, mobile phone, pen, plastic bottle, plush doll, TV controller, scissors, socks*, and *towel*. There are two types of image: Source (i.e., handheld object images) and Target (i.e., home explored images). In the target domain, we consider the unknown class not included in the source dataset. We describe the details of the source and target data in the following.

Source (handheld object images): Figure 2 shows some examples of source images. We show each object to the HSR in front of a uniform background while rotating it by hand. Therefore, each image contains exactly one object. We manually annotate both the class label and object bounding box for each image. There are 53,991 images for training and 16,000 images for testing, which are denoted as $\text{Source}_{\text{train}}$ and $\text{Source}_{\text{test}}$, respectively. Although $\text{Source}_{\text{test}}$ does not directly correspond to the goal of the COSDA-HR dataset, it can be useful for the comparative evaluation of the methods in within-domain scenarios.

Target (home explored images): We manipulate the HSR and collect images in an experimental home environment as illustrated in Figure 3 (bottom). In the environment, multiple objects are manually placed at plausible locations such as the table and kitchen. The HSR starts exploring from the red circle and finishes the exploration at the green circle going along the path indicated by the black arrow. We conduct 22 explorations randomly changing the objects and their locations for each exploration. Half of the explorations were conducted in the daytime with sufficient sunlight while the other half of the explorations were conducted at night where only room lights are available. The examples of images in the target data are shown in Figure 3 (top). Out of the 22 explorations, 10 are used for training (i.e., domain adaptation), another 10 are used for testing, and the last two are used for tuning an objectness detector. They are denoted as $\text{Target}_{\text{train}}$, $\text{Target}_{\text{test}}$, and $\text{Target}_{\text{tune}}$, respectively. While $\text{Target}_{\text{train}}$ does not have any annotation, $\text{Target}_{\text{test}}$ has ground-truth object label and bounding box annotations, which are only used for evaluation purposes. $\text{Target}_{\text{tune}}$ only has object bounding boxes (without class labels).

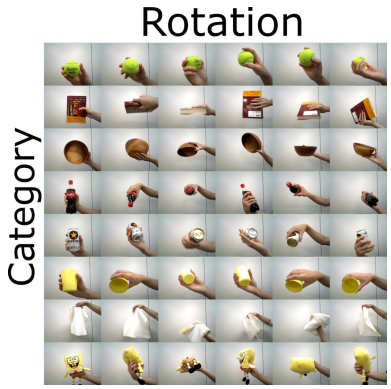

Figure 2: Examples of source images. Each object belongs to a different super category and is shown to the robot while being rotated by hand.

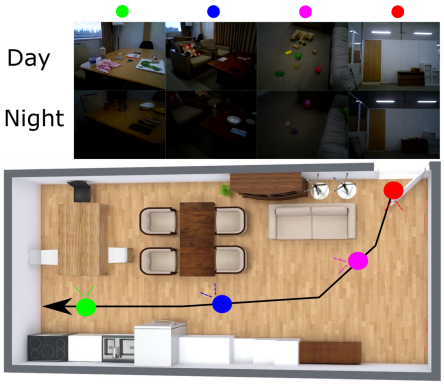

Figure 3: Experimental home environment. The bottom image shows the map and the trajectory of the robot. Each colored circle indicates the position of the robot, and the corresponding target images captured there are shown on the top.

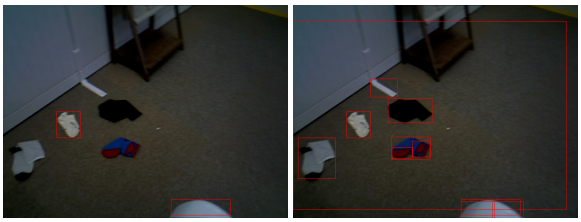

Figure 5: Failure case of object proposals by object detector with default hyperparameters. The left and right images are object proposals with default and tuned hyperparameters, respectively.

3.2. Converting to a Classification Problem

Unlike the source domain dataset, target domain images are not object-centric, and thus we should somehow detect the object regions both in the training phase (i.e., domain adaptation using $\text{Target}_{\text{train}}$) and in the testing phase. Because the object detection part will significantly affect the overall outcome of a system but is not the central interest of the COSDA-HR dataset, we also provide a preprocessed target dataset by fixing the detection process to focus on the problem of the COSDA-HR dataset. Specifically, we crop the object images with an off-the-shelf objectness detector and set up the problem as the standard classification. To do this, we use a general objectness detector that is implemented with the RetinaNet [20] pre-trained on COCO dataset [21]. As can be seen in Figure 5, we found that the pre-trained model often fails to detect target objects in our environment, we tune the hyperparameter (i.e., detection threshold) of the detector with $\text{Target}_{\text{tune}}$ data so that it maximizes the recall score[2] for the tuning data because this is the very first step in the detection pipeline and it is more important to keep as many true positives as possible. As a result, our objectness detector achieves 0.015 precision and 0.967 recall on $\text{Target}_{\text{tune}}$ even though 7 super categories in the COSDA-HR dataset does not exist in categories of COCO dataset.

We apply the objectness detector to $\text{Target}_{\text{train}}$ and obtain the set of cropped object images, which is denoted as $\text{Target}^{*}_{\text{train}}$. It has 205,392 images in total and the minimum, average, and maximum size across all cropped images are 24 x 15, 100 x 140, and 480 x 640 pixels, respectively. It is a very noisy open-set dataset where most of the data belong to the unknown class that contains both unknown objects and false-positive object proposals. Some examples can be seen in Figure 1 (in green boxes on the target side).

As for the testing data, we apply the objectness detector and crop the regions whose IoU with any one of the original ground-truth bounding boxes is larger than 0.5, which are regarded as the object regions for the corresponding labels. If a proposed region does not have an IoU of more than 0.5 with any ground-truth boxes, it is regarded as an unknown class. In this way, we set up a classification dataset by sampling 15,094 images for the known 160 classes (i.e., closed set) and 15,094 images for the unknown class, which we name $\text{Target}^{*}_{\text{test}}$. The correct classification of those images means the correct object detection (with unknown rejection) in the overall pipeline.

Table 2 summarizes the properties of the original and the converted datasets. In the rest of this paper, we focus on the classification problem with $\text{Source}_{\text{train}}$, $\text{Target}^{*}_{\text{train}}$, and $\text{Target}^{*}_{\text{test}}$

[2]The bounding box predictions are regarded as correct if its IoU with any one of the ground-truth bounding boxes is greater than 0.5. Note that we do not fine-tune the detection model itself.

Subsets	# Explorations (day/night)	# Images (known/unknown)	# Classes	Annotations
$\text{Source}_{\text{train}}$	-	53,991 (53,991/0)	160	bbox + label
$\text{Source}_{\text{test}}$	-	16,000 (16,000/0)	160	bbox + label
$\text{Target}_{\text{train}}$	10 (5/5)	1,552	160+1	
$\text{Target}_{\text{test}}$	10 (5/5)	7,851	160+1	bbox + label
$\text{Target}_{\text{tune}}$	2 (1/1)	1,390	160+1	bbox
$\text{Target}^{*}_{\text{train}}$	10(5/5)	205,392	160+1	
$\text{Target}^{*}_{\text{test}}$	10(5/5)	30,188 (15,094/15,094)	160+1	label

Table 2: Properties of the COSDA-HR dataset. In addition to the domain shift, the target data include the unknown class that contains unknown objects not presented in the source. Target* represents the pre-processed dataset of Target.

3.3. Evaluation

Clearly, there is a trade-off between the performances of closed-set accuracy and unknown rejection. Considering that there could be an infinite amount of unknown objects in a real home environment, we believe it is more important to be able to correctly reject a majority of unknown inputs in the first place. The standard criteria, such as mean class accuracy and macro F1 score, can not be used to properly evaluate this point. Therefore, we propose the t-threshold top-m accuracy and denote it as top-m (t=n). It represents the top-m accuracy in the closed set classes when the accuracy (recall) for the unknown class is larger than n. In our experiments, we mainly evaluate the recognition ability when m and n are 1 and 0.9, respectively.

4. Experiments

In the following experiments, except for the simple open-set recognition methods, we train the model with $\text{Source}_{\text{train}}$ + $\text{Target}^{*}_{\text{train}}$ and evaluate with $\text{Target}^{*}_{\text{test}}$. For open-set methods, we only use $\text{Source}_{\text{train}}$ for training as they have no clear way of utilizing $\text{Target}^{*}_{\text{train}}$ data. In addition to the open-set evaluation of top-1 (t=0.9), we also evaluate top-1 (t=0.0), which is the case when an unknown class is not rejected and corresponds to the standard closed-set evaluation of top-1 accuracy within the known 160 classes.

We use ResNet-18 [10] pre-trained on ImageNet-1k [33] as the backbone feature extractor for all experiments. Image preprocessing and data augmentation are applied in the same way that are shown in [10]. Other implementation details (e.g., optimizer, learning schedule) are shown in the supplementary material.

4.1. Open Set Domain Adaptation

We first consider the open set domain adaptation without the continual learning issue in which we use the whole training data in one step. The performances of various methods are summarized in Table 3. First, we test some state-of-the-art unsupervised domain adaptation methods (i.e., AFN [46], BSP [6], DAAN [49], DTA [18] and MRAN [54]), which do not consider the open-set problem by themselves. As they cannot handle the unknown class, we only evaluate the closed top-1 (t=0.0) accuracy for these methods. As can be seen in Table 3, MRAN [54] achieves the best performances of 37.35%, but this is relatively low even though we ignore the unknown samples in testing. This result indicates the difficulty of domain alignment with the existence of unknown class data in $\text{Target}_{\text{train}}$.

Then, we consider the open-set domain adaptation problem. Not surprisingly, simple open-set recognition methods do not work because of the huge domain shift. A straightforward approach to OSDA is to combine the methods of open-set recognition and domain adaptation individually developed in each field. For this category, MRAN + Softmax thresholding achieves 13.38% top-1 (t=0.9) accuracy.

A more plausible approach is to use the methods specially designed for open-set domain adaptation, which consider both problems simultaneously. We test two standard methods in this category: OPDABP [36] and UDA [48]. Overall, the best score is achieved by UDA.

4.2. Continual Open Set Domain Adaptation

On the basis of the results discussed in the previous section, we consider the open-set domain adaptation with continual learning as our final goal. In this experiment, 160 classes are randomly split into 10 subsets to form the sequence of continual learning. The performance is evaluated in terms of the final 160 class classification after inputting the 10 subsets one by one. We check the performance using four types of criteria on top-1/10 (t=0.0), and top-1/10 (t=0.9).

We combine state-of-the-art continual learning methods with open-set domain adaptation methods evaluated in the previous section. For the continual learning methods, we use LwF [19] as a representative method without memory, and the simple rehearsal method (Rehearsal) and BiC [50] as memory-based methods because they achieved the best results among several methods in our preliminary experiment using the closed dataset in the source domain (see supplementary material). We set the size of the memory as 2000 for the memory-based methods. "Cumulative" corresponds to the standard off-line training strategy and can be

Method	DA	OS	top-1(t=0.0) (%)	top-1(t=0.9) (%)
AFN [46]	✓		22.63	NA
BSP [6]	✓		29.80	NA
DAAN [49]	✓		31.33	NA
DTA [18]	✓		31.21	NA
MRAN [54]	✓		37.35	NA
Softmax thresholding		✓	15.63	4.03
Openmax [3]		✓	10.58	3.00
CROSR [47]		✓	6.80	2.29
OPDABP [36]	✓	✓	19.04	6.73
UDA [48]	✓	✓	35.53	**15.89**
MRAN+Softmax	✓	✓	**37.35**	13.38
MRAN+Openmax	✓	✓	17.66	8.91

Table 3: Closed-set (t=0.0) and open-set (t=0.9) evaluation results on the COSDA-HR dataset in open-set domain adaptation setting. DA and OS indicate that a method covers domain adaptation and open-set issues, respectively.

Method	top-1 (t=0.0) (%)	top-1 (t=0.9) (%)	top-10 (t=0.0) (%)	top-10 (t=0.9) (%)
LwF+UDA	6.85	2.93	24.96	8.68
Rehearsal+UDA	32.75	16.71	72.61	**29.03**
BiC+UDA	32.39	**17.27**	72.08	28.84
UDA (Cumulative)	**34.38**	15.89	**73.52**	24.22
BiC+MRAN+Softmax	31.93	15.17	71.47	15.18
BiC+MRAN+Openmax	11.24	0.90	42.23	2.89

Table 4: Closed-set (t=0.0) and open-set (t=0.9) evaluation results on the COSDA-HR dataset in continual open-set domain adaptation setting.

regarded as the upper-bound baseline for continual learning.

The results are shown in Table 4. It shows that while a no-memory method (LwF + UDA) performs significantly poorer than the non-continual baseline (UDA(Cumulative)), memory-based continual learning methods can somewhat alleviate the problem of catastrophic forgetting. The most interesting finding is that while UDA(Cumulative) outperforms its continual versions (w/ Rehearsal and BiC) on closed-set evaluation (t=0.0), which is not surprising, continual ones perform better in terms of open-set evaluation (t=0.9). This is perhaps because memory-based CL methods retain only a limited amount of informative examples that lead to loose estimation of the posterior probability distributions of known labels. This result indicates the importance of considering continual learning effects simultaneously with open-set and domain adaptation to address the problems in the COSDA-HR dataset.

Overall, the best combination in this experiment is BiC + UDA, which achieves 17.27% in top-1 (t=0.9), but we should say that the performances of those baseline methods are far from satisfactory for home robots. In addition to the significant domain shift, the open-set issue is hardly handled in these methods. For example, BiC + UDA achieves 32.39% on closed-set evaluation top-1 (t=0.0) but only achieves 17.27% open-set top-1 (t=0.9). It means that almost half of the correct prediction is wrongly detected as the unknown class and rejected. More studies are required to alleviate these difficulties. We also provide some empirical analyses of hard examples in the supplementary material.

5. Analysis

5.1. Do better ImageNet-1k classifiers generalize better?

It is empirically shown that the better the performance in ImageNet-k, the better the transferability to other recognition tasks [15]. The results in Tables 3 and 4 are obtained using ResNet-18 as the backbone feature extractor. In this section, we investigate how much a strong backbone contributes to improving performance.

The relationship between top1-error in ImageNet-1k and open-set (t=0.9) evaluation results in the COSDA-HR dataset with continual open-set domain adaptation setting is shown in Figure 6. There is a trend that a stronger backbone gives better performance in the COSDA-HR dataset. The result is still far from satisfactory for home robots even when some strong backbones are used. In addition, some backbones (i.e., ResNet50, WideResNet50 [51], WideResNet101, DenseNet121 [11] and ResNeXt50 [45])

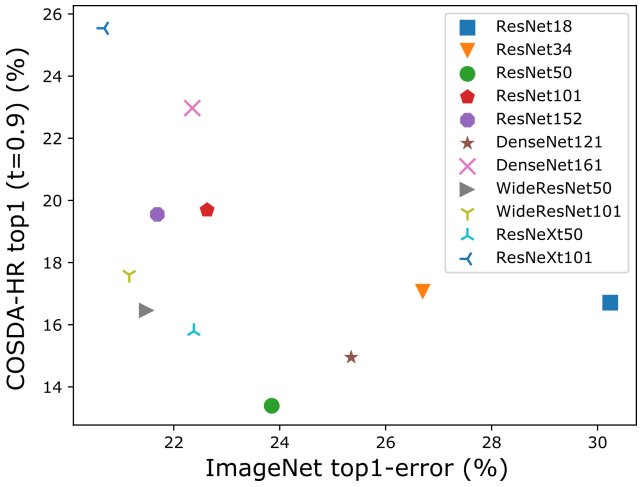

Figure 6: Relationship between top1-error in ImageNet-1k and open-set (t=0.9) evaluation results in the COSDA-HR dataset with continual open-set domain adaptation setting.

Figure 7: Examples of target images in the COSDA-HR dataset under different light conditions.

Condition	top-1 (t=0.9) (%)
Day	17.85
Night	15.95

Table 5: Open-set (t=0.9) evaluation results on the COSDA-HR dataset under different light conditions.

show weaker performances than ResNet18 in the COSDA-HR dataset even though ResNet18 shows the weaker performance in ImageNet-1k. It implies that strong backbones may give better performance, however, it is not surely guaranteed. We consider that using a strong backbone can not be the solution to the COSDA-HR dataset.

5.2. Empirical analysis in hard examples

We analyze how our baseline makes a mistake in the COSDA-HR dataset.

Light condition. $\text{Target}^*_{\text{test}}$ has two types of example that are taken during the day and at night. We examine how such light conditions affect recognition ability. As can be seen in Figure 7, there are two types of light condition in the COSDA-HR dataset: day (i.e., sunlight + indoor lights) and night (i.e, only indoor lights). The results of the performance under each light condition are shown in Table 5. It shows that the recognition ability is worse at night than during the day even though the examples under both conditions appear similar.

Super category	top-1 (t=0.9) (%)	False rejection rate (%)
glass bottle	45.38	28.53
socks	39.23	44.09
book	26.77	50.0
scissors	0.0	89.62
dish	4.31	80.43
plastic bottle	7.07	55.31

Table 6: Three easiest and hardest super categories. The top three and bottom three are the result of easiest and hardest categories, respectively.

Misclassified examples. There are 16 super category and we investigate which super categories are difficult to recognize. Table 6 shows three super categories that achieve the highest top-1 (t=0.9) and the lowest ones. False rejection rate represents how many examples are rejected as the unknown class. It shows that some super categories are mostly rejected as the unknown class (e.g., scissors and dish). There is the trend that the less the rejection rate in the super category, the better the performance. It implies that performance should be much improved by developing better rejection mechanisms.

6. Conclusion

In this paper, we proposed the COSDA-HR dataset to measure the ability of home robots to recognize naturally arranged objects in a room by training with handheld objects. It is built to help to develop a user-friendly object teaching system for home robots. Since we consider a realistic situation for home robot applications, the COSDA-HR dataset requires dealing with three problems of domain adaptation, open-set recognition, and continual learning simultaneously.

We also performed experiments on the COSDA-HR dataset by thoroughly combining state-of-the-art methods of each of the three problems and confirmed that simple combinations still provide far from satisfactory results. Much work is required to mitigate the difficulties by dealing with the three problems simultaneously instead of solving each problem separately.

Although we separately used an objectness detector to focus on the classification problem in this paper, it would be interesting to develop an end-to-end learning method including the object detection process. It is also promising to integrate the depth information to improve the performance. We would like to leave these issues as our future work.

References

[1] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *NIPS*, 2019.

[2] Mahsa Baktashmotlagh, Masoud Faraki, Tom Drummond, and Mathieu Salzmann. Learning factorized representations for open-set domain adaptation. *ICLR*, 2019.

[3] Abhijit Bendale and Terrance Boult. Towards open set deep networks. *CVPR*, 2016.

[4] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. *CVPR*, 2017.

[5] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *ICLR*, 2019.

[6] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. *ICML*, 2019.

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[8] Zongyuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *BMVC*, 2017.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NIPS*, 2014.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[11] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. *CVPR*, 2017.

[12] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. *ACL*, 2007.

[13] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. *CVPR*, 2019.

[14] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[15] Simon Kornblith, Jon Shlens, and Quoc V. Le. Do better imagenet models transfer better? *CVPR*, 2019.

[16] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, Rahul M. V., and R. Venkatesh Babu. Towards inheritable models for open-set domain adaptation. *CVPR*, 2020.

[17] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. *ICRA*, 2011.

[18] Seungmin Lee, Dongwan Kim, Namil Kim, and Seong-Gyun Jeong. Drop to adapt: Learning discriminative features for unsupervised domain adaptation. *ICCV*, 2019.

[19] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection (best student paper award). *ICCV*, 2017.

[21] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *CVPR*, 2014.

[22] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. *CVPR*, 2019.

[23] Mohammad Reza Loghmani, Barbara Caputo, and Markus Vincze. Recognizing objects in-the-wild: Where do we stand? *ICRA*, 2018.

[24] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. *Proceedings of the 1st Annual Conference on Robot Learning*, 78, 2017.

[25] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv*, 2015.

[26] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *NIPS*, 2018.

[27] David Lopez-Paz and Marc Aurelio Ranzato. Gradient episodic memory for continual learning. *NIPS*, 2017.

[28] Yingwei Pan, Ting Yao, Yehao Li, Chong-Wah Ngo, and Tao Mei. Exploring category-agnostic clusters for open-set domain adaptation. *CVPR*, 2020.

[29] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. *ICCV*, 2017.

[30] G. Pasquale, C. Ciliberto, L. Rosasco, and L. Natale. Object identification from few examples by improving the invariance of a deep convolutional neural network. *IROS*, 2016.

[31] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. *AAAI*, 2018.

[32] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. *CVPR*, 2017.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[34] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. *ECCV*, 2010.

[35] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. *CVPR*, 2018.

[36] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. *ECCV*, 2018.

[37] Qi She, Fan Feng, Xinyue Hao, Qihan Yang, Chuanlin Lan, Vincenzo Lomonaco, Xuesong Shi, Zhengwei Wang,

Yao Guo, Yimin Zhang, Fei Qiao, and Rosa H. M. Chan. Openloris-object: A dataset and benchmark towards lifelong object recognition. *arXiv*, 2019.

[38] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. *AAAI*, 2018.

[39] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *NIPS*, 2017.

[40] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. *ECCV*, 2016.

[41] Dan Teng and Sakyasingha Dasgupta. Continual learning via online leverage score sampling. *arXiv*, 2019.

[42] TOYOTA MOTOR CORPORATION. *Toyota shifts home helper robot R&D into high gear with new developer community and upgraded prototype*. `https://global.toyota/en/detail/8709541`. Accessed 7 Nov, 2020.

[43] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *CVPR*, 2017.

[44] Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. *arXiv*, 2019.

[45] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *CVPR*, 2017.

[46] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. *ICCV*, 2019.

[47] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. *CVPR*, 2019.

[48] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Universal domain adaptation. *CVPR*, 2019.

[49] Chaohui Yu, Jindong Wang, Yiqiang Chen, and Meiyu Huang. Transfer learning with dynamic adversarial adaptation network. *arXiv*, 2019.

[50] Lijuan Wang Yuancheng Ye Zicheng Liu Yandong Guo Yun Fu Yue Wu, Yinpeng Chen. Large scale incremental learning. *arXiv*, 2019.

[51] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *BMVC*, 2016.

[52] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv*, 2017.

[53] Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. On learning invariant representation for domain adaptation. *arXiv*, 2019.

[54] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Jingwu Chen, Zhiping Shi, Wenjuan Wu, and Qing He. Multi-representation adaptation network for cross-domain image classification. *Neural Networks*, 119:214–221, 2019.