

Contrastive Learning for Cross-Domain Open World Recognition

Original

Contrastive Learning for Cross-Domain Open World Recognition / Cappio Borlino, Francesco; Bucci, Silvia; Tommasi, Tatiana. - ELETTRONICO. - (2022), pp. 10133-10140. (Intervento presentato al convegno 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022) tenutosi a Kyoto (Giappone) nel 23-27 ottobre 2022) [10.1109/IROS47612.2022.9981592].

Availability:

This version is available at: 11583/2971072 since: 2022-09-07T14:29:54Z

Publisher:

IEEE

Published

DOI:10.1109/IROS47612.2022.9981592

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

IEEE postprint/Author's Accepted Manuscript

©2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

Contrastive Learning for Cross-Domain Open World Recognition

Francesco Cappio Borlino¹, Silvia Bucci¹, and Tatiana Tommasi¹

Abstract—The ability to evolve is fundamental for any valuable autonomous agent whose knowledge cannot remain limited to that injected by the manufacturer. Consider for example a home assistant robot: it should be able to incrementally learn new object categories when requested, but also to recognize the same objects in different environments (rooms) and poses (hand-held/on the floor/above furniture), while rejecting unknown ones. Despite its importance, this scenario has started to raise interest in the robotic community only recently and the related research is still in its infancy, with existing experimental testbeds but no tailored methods. With this work, we propose the first learning approach that deals with all the previously mentioned challenges at once by exploiting a single contrastive objective. We show how it learns a feature space perfectly suitable to incrementally include new classes and is able to capture knowledge which generalizes across a variety of visual domains. Our method is endowed with a tailored effective stopping criterion for each learning episode and exploits a self-paced thresholding strategy that provides the classifier with a reliable rejection option. Both these novel contributions are based on the observation of the data statistics and do not need manual tuning. An extensive experimental analysis confirms the effectiveness of the proposed approach in establishing the new state-of-the-art. The code is available at https://github.com/FrancescoCappio/Contrastive_Open_World.

I. INTRODUCTION

Trustworthy robotic assistants for industrial, home and street environments should be able to recognize multiple objects and detect new unseen categories. Although we are taking as example scenarios that offer different levels of control, all of them share the need for robust visual systems: they should generalize to a variety of real-world target conditions (*e.g.* change in viewpoint, camera equipment, illumination, home, weather, country) while maintaining the ability to identify any unknown object and possibly learn its category over time. For a practical example let’s consider a home assistant robot. It cannot be limited to recognize only a pre-defined set of classes for which it was programmed: it should be able to incrementally learn new objects when its owner handles them and then to recognize those objects when they are naturally arranged in different rooms, without getting confused by others for which it has not received instructions [1]. Managing all these tasks at once is extremely challenging. Most of the existing successful deep learning models are applied on simplified problems under *closed-set* and *closed-domain* conditions. The former considers a match between the training and test category sets so that the











Setting	Learn	Predict
Open Set Recognition (OS)		
Class Incremental Learning (CIL)		
Open World Recognition (OWR)		
Domain Generalization (DG)		
Cross-Domain Open World Recognition (CD-OWR)		

Fig. 1. Cross-Domain Open World Recognition (CD-OWR) and other related settings compared with respect to the training and deployment conditions. Each shape indicates a different category. The sketchy style specifies a different data distribution with respect to that of the white empty shapes. The task index higher than 0 refers to the settings where multiple incremental learning steps are needed.

classes available while learning are assumed to be the only ones that could be ever encountered at deployment time. The latter means neglecting situations in which train and test samples have the same semantic content but come from different visual distributions (also known as domains).

Several learning frameworks have been defined to push the boundaries of object recognition towards more realistic *open-world* scenarios (see Figure 1). *Open-Set recognition* (OS) deals with the identification of novel classes at test time that were not present in the training phase, while also maintaining the recognition performance for known classes [2], [3]. *Class Incremental Learning* (CIL) focuses on extending an original model to accommodate novel classes in subsequent incremental tasks [4], [5], [6]. Existing works on *Open World Recognition* (OWR) combine OS and CIL but mainly disregard domain shift conditions [7], [8], [9]. Indeed, a change in domain between training and test data can create confusion in the identification of the novel categories, and consequently, make their inclusion in the training process even more challenging.

Cross-Domain Learning (CD) aims at improving the performance of a model trained on a labeled source domain when tested on data of a distinct target domain [10], [11]. Several *Domain Generalization* (DG) and *Domain Adapta-*

¹F. Cappio Borlino, S. Bucci and T. Tommasi are with the DAUIN Department at Politecnico di Torino, Italy. F. Cappio Borlino and T. Tommasi are also affiliated with the Italian Institute of Technology, Italy. {francesco.cappio, silvia.bucci, tatiana.tommasi}@polito.it

TABLE I

COMPARISON WITH EXISTING OWR (OS+CIL), DG AND CIL APPROACHES. HPS INDICATE THE HYPERPARAMETERS.

Method	No. of Losses	No. of HPs	Open-Set Recognition	Domain Generalization	Class Incremental Learning
NNO [14]	1	1	✓		✓
DeepNNO [8]	2	1	✓		✓
B-DOC [9]	3	2	✓		✓
SS-IL [6]	2	0			✓
RR [15]	2	1		✓	
SC [16]	1	1		✓	
RSDA [17]	2	1		✓	
SagNet [18]	3	2		✓	
COW	1	2	✓	✓	✓

tion (DA) approaches have been developed for this purpose, with DG methods working without accessing the unlabeled target data at training time. An extension of DG to the open-set scenario was recently presented in [12], but the proposed method strongly relies on the availability of multiple data sources. There have been also some attempts to define a wider *Cross-Domain Open-World Recognition* (CD-OWR) setting with works mainly focusing on benchmarks and naïve combinations of existing methods (OWR+DA [1], OWR+DG [13]): they have highlighted how difficult the setting is, remaining far from solving it.

To the best of our knowledge, we propose the first approach that deals with all the challenges of the CD-OWR scenario at once. We show how a single supervised contrastive objective is suitable for open world recognition while also promoting domain generalization. Specifically, in the hyperspherical feature space obtained via contrastive learning, samples of the same class tend to cluster together regardless of their domain, while novel categories appear in low-density regions (see Fig. 2). By considering the Nearest Class Mean [7] logic which is the basis of many OWR methods, it becomes clear that the described embedding is an ideal environment where a simple rejection rule can be applied on sample-to-prototype distances to identify novel categories. Moreover, our approach does not need class-specific rejection thresholds as the learned feature space pushes all clusters to have similar structures and distances, further simplifying the task. Our key contributions are the following:

- We propose our *Contrastive Open-World* (COW) approach: it is a straightforward method (see Table I) able to deal with CD-OWR by simply exploiting the highly structured feature space obtained via contrastive learning.
- COW manages incremental class learning by using a tailored stopping criterion at each learning episode. Its implementation provides empirical guarantees on the quality of the model learned after each incremental step.
- COW exploits a novel thresholding strategy free from manual tuning which effectively separates known and unknown target data. The decision adapts to the peculiarities of the considered datasets since it is based on the observed sample distribution in the learned feature space.
- A thorough experimental analysis of existing CD-OWR benchmarks confirms the effectiveness of COW which sets the new state-of-the-art.

II. RELATED WORK

A. Open World Recognition

Hard-coded recognition skills are clearly not enough for autonomous robots that operate in unconstrained environments. *Class Incremental* (or *Continual*) *Learning* provides support with strategies that update pretrained models by including new classes once they are observed. In order to be effective, this should be done with as little access to the old data as possible, while also avoiding to forget previous knowledge. One of the early incremental approaches was based on the Nearest Class Mean (NCM, [7]) classifier, which computes the mean of the feature vectors for the training samples and each test sample is assigned to the nearest class prototype. In the more recent literature, there are two main types of approaches. Some methods keep a small memory buffer of previous data in order to replay it while learning new classes [4], [19], [6]. Other more complex approaches do not require memory, but exploit extra distillation objectives, constraints on network weights, or rely on generative solutions to reproduce data of previously seen classes [5], [20]. Robotics applications for methods of the first family were presented in [21], [22], [23]. The growing interest in this field is also testified by the ever wider range of datasets designed for it [1], [24], [25].

Dealing with unconstrained learning conditions means also not knowing a priori which classes will be encountered at test time. *Open Set Recognition* approaches are able to distinguish such unknown objects from the known ones and have been extensively studied both by the computer vision and robotics communities [2], [3], [26]

Finally, *Open World Recognition* combines the two previous settings. OWR was first introduced in [14] which proposed NNO, a simple extension of the standard NCM strategy including an unknown rejection policy. Despite its significance for real applications, few other works followed the mentioned seminal paper: DeepNNO [8], the deep version of NNO, and B-DOC [9] that includes clustering objectives and class-specific rejection thresholds.

B. Domain-Shift

The discrepancy between training and test data is one of the most relevant problems in robotics where many processes are designed completely in simulated environments, but then need to be applied in the real world [27], [28], [29], [30]. Neglecting the domain shift between source (labeled training) and target (unlabeled test) data leads to brittle methods and frequent fails. In *Domain Adaptation* the target is available during training and is used to adapt the source model [10]. Some online variants [31] take advantage of a stream of incoming target samples at training time: in this case, there is less available target information and the adaptation process becomes slower [32], [33]. In *Domain Generalization* the target domain is not known while training on the source: this scenario requires to build models that are robust to any deployment conditions. This goal is obtained by an effective exploitation of available source data, often

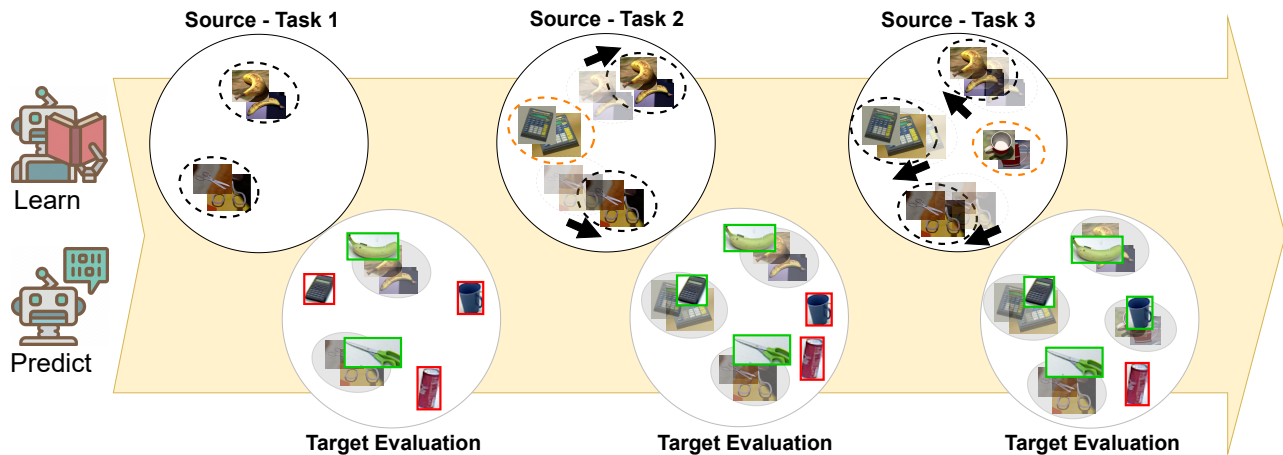


Fig. 2. Overview of COW. During training, the classes are incrementally learned at each time step and the old clusters move on the hypersphere to make room for new semantic categories. During the evaluation the target is mapped on the trained hypersphere: the samples far from any existing centroid are marked as unknown (red square), and the samples near enough to a centroid are classified as known with the category of the nearest one (green square).

considering to have access to multiple different source domains. Most DA and DG works consider the exact same class set shared by source and target, while more recently open-set DA was studied in [34], [35], [36], [37]. Only one work focused on open-set DG under the assumption of multiple available source domains [12]. Overall the main existing adaptive approaches can be organized in few groups: discrepancy-based and feature alignment methods [38], [39], adversarial learning [40], data augmentation techniques [17], [18], [41], meta-learning [42], and self-supervised learning based approaches [15], [43].

C. Contrastive Learning

Self-supervised Contrastive Learning aims at maximizing the agreement among multiple augmentations of the same sample while pushing different instances far apart. This strategy has been recently implemented with different variants [44], [45], [46]: despite not relying on any annotation, the learned hyperspherical embedding (L2 normalized features) captures reliable semantic information, as demonstrated by the effectiveness of the contrastive models used as pretraining for various downstream tasks [47]. The merits of contrastive learning become even more evident when considering its supervised formulation [48], simply obtained by exploiting class labels to build positive and negative sample pairs. Samples of the same class are encouraged to cluster together regardless of their intra-class appearance variation, with also large inter-class distances. As a consequence, samples belonging to novel classes end up in low-density regions and measuring the distance to the closer class prototype provides useful rejection information. This approach has been used in several tasks going from novelty detection [49] to cross-domain generalization [50], open-set domain adaptation [37], and class incremental learning [51].

Previous work has only scratched the surface of the challenging problem of Cross-Domain Open World Recognition: existing DA and DG approaches have been applied to open world settings to evaluate their potentialities and limits [1],

[13]. In this work, we design instead a tailored method for CD-OWR by leveraging supervised contrastive learning.

III. METHOD

A. Notations and problem setting

Let us start from the initial training set $S_0 = \{\mathbf{x}_i, y_i\}_{i=0}^{N_{S_0}}$ where $\mathbf{x}_i \in \mathbb{X}$ is the image and y_i the corresponding category label from the label set Y_{S_0} . We assume to subsequently receive new incremental tasks $\{S_1, \dots, S_K\}$ such that their label sets don't overlap *i.e.* $Y_{S_k} \cap Y_{S_{k'}} = \emptyset \forall k, k' \in [0, \dots, K]$ and $k \neq k'$. We call *source domain* the entire labeled training set $S = \{\bigcup_{k=0}^K S_k\}$ that is drawn from data distribution p_{sr} . The unlabeled test set $T = \{\mathbf{x}_i\}_{i=0}^{N_T}$ is not seen during training and is drawn from the target distribution p_{tg} , with $p_{sr} \neq p_{tg}$. Moreover, the target contains both known and unknown categories. After training on each source task, the goal is to predict for a target test sample whether it is from one of the learned categories or it is unknown. More precisely, our goal is to train a function $f: \mathbb{X} \rightarrow \{Y_s \cup u\}$ such that the target sample is mapped either to one of the semantic categories learned until the current task $Y_s = \{\bigcup_{k=0}^t Y_{S_k}\}$ or to the unknown class u . We consider f made by three components: a feature extractor $g: \mathbb{X} \rightarrow \mathbb{Z}$ that maps each image in the feature space, a scoring function $\eta: \mathbb{Z} \rightarrow \mathbb{R}^{|Y_s|}$ that maps the features to a score vector representing the probability that the sample is from one of the learned categories, and finally $\omega: \mathbb{R}^{|Y_s|} \rightarrow \{Y_s \cup u\}$ that will make the final prediction.

B. Supervised Contrastive Learning

We propose to train a contrastive model to obtain a highly structured feature space \mathbb{Z} ready to accommodate the new categories that sequentially arrive with each task. The self-supervised contrastive learning loss [44], [45] models the feature space by maximizing the similarity between each instance and its augmented version (*positive pair*), while minimizing the similarity between two different instances (*negative pair*). In particular, for each training sample $\{\mathbf{x}_j, y_j\}$, an augmented version $\{\mathbf{x}'_j, y_j\}$ is produced through standard transformations (*e.g.* grayscale, random crop, color

jittering), doubling the original batch $B = \{j = 1, \dots, 2J\}$. In our setting we have the category labels of training data, hence we use a *supervised* version of the contrastive learning approach [48], which simply modifies the training objective by exploiting sample labels to create positive (same class) and negative (different classes) pairs.

We consider the feature extractor g composed of an Encoder E and a Projection head P . For each sample in B , we obtain the representation $\mathbf{z}_j = g(\mathbf{x}_j) = E(P(\mathbf{x}_j))$. The final learning objective that we use for the training is:

$$\mathcal{L}_{SupCtr} = \sum_{j=1}^{2J} \frac{-1}{|\pi(j)|} \sum_{j' \in \pi(j)} \log \frac{\exp(\sigma(\mathbf{z}_j, \mathbf{z}_{j'})/\tau_e)}{\sum_{n \in v(j)} \exp(\sigma(\mathbf{z}_j, \mathbf{z}_n)/\tau_e)}. \quad (1)$$

Here $v(j) = B \setminus \{j\}$ is the double batch without the *anchor* sample of index j , and $\pi(j) = \{j' \in v(j) : y_{j'} = y_j\}$ is the set of all the positive pairs. Finally $\tau_e \in \mathbb{R}^+$ is a temperature parameter, and $\sigma(\cdot, \cdot)$ is the cosine similarity. Thanks to the L2-normalization of the cosine similarity we learn features that lay on a hyperspherical surface and we keep this embedding geometry for all the steps of our learning procedure.

C. Feature space structure and statistics

The objective described above pushes the data to form compact and well-separated class clusters on the surface of the hypersphere [52]. This structure allows to easily compute some statistics about data distribution, as done in [37]. First of all we define the prototype of each known class $y_s \in Y_s$ by computing the corresponding feature average $\mathbf{h}_{y_s} = \frac{1}{|y_s|} \sum_{k \in y_s} \mathbf{z}_k$, re-projected on the unit hypersphere. Here $|y_s|$ indicates the number of samples of class y_s and k is the index that runs on all them. We also define the angular distance measure $d_a(\mathbf{z}_i, \mathbf{z}_j) = \{1 - \sigma_{[0,1]}(\mathbf{z}_i, \mathbf{z}_j)\}$ which rescale the cosine similarity in $[0, 1]$ and translates it. We are interested in two feature space statistics:

- the *class sparsity*, which measures the average distance between a prototype \mathbf{h}_{y_s} and the nearest one among the others \mathbf{h}_* : $\theta = \frac{1}{|Y_s|} \sum_{y_s \in Y_s} d_a(\mathbf{h}_*, \mathbf{h}_{y_s})$;
- the *class compactness*, which measures the average distance between training samples and the corresponding class prototypes: $\phi = \frac{1}{|Y_s|} \sum_{y_s \in Y_s} \left\{ \frac{1}{|y_s|} \sum_{k \in y_s} d_a(\mathbf{h}_{y_s}, \mathbf{z}_k) \right\}$

These two metrics describe the data distribution and can be used as reference to make decisions both on the stopping criterion for each learning episode, and on the known-unknown class separation.

D. Incremental protocol

To keep the focus on the effectiveness of the supervised contrastive loss function for the CD-OWR task, we do not implement any complex additional module to avoid forgetting in the incremental procedure. We, adopt only two simple strategies: i) as done by other incremental learning algorithms we keep a (limited- and fixed-size) replay buffer containing a subselection of samples from previous tasks; ii) we perform class balancing at the batch level, by putting in each training mini-batch at least two samples of each class.

We expect our learning objective to manage the data and progressively make room on the hyperspherical feature space to accommodate new classes while exploiting replay samples to maintain reserved space for the old ones.

E. Stop-training criterion

Intuitively class clusters in our feature space cannot be well separated if $\theta < 2\phi$. In fact, we can see ϕ as a measure of the radius of clusters: if the distance between two class centroids is lower than the sum of their radii the two clusters will inevitably overlap. In this case, samples of the two classes cannot be distinguished. Moreover, with no *empty space* between class clusters, there's no space for unknown data either. In order to avoid this condition, we can impose a constraint on the quality of the feature space for our output model. Given that the compactness and separation of the clusters increase during training, we can enforce the described relation between θ and ϕ by using a specific stopping criterion in the learning procedure. We consider each learning task as converged only when

$$\lambda > 1 + \varepsilon, \quad \text{with } \lambda = \frac{\theta}{2\phi} \quad \text{and } \varepsilon \geq 0. \quad (2)$$

Here ε can be seen as a *minimum desired margin* between two clusters.

F. Threshold definition

By following the same logic of the NCM strategy, our scoring function is based on the distance of each target sample to its nearest source class prototype: $\eta(\mathbf{z}^t) = d_a(\mathbf{h}_{y_s}, \mathbf{z}^t)$. This approach is particularly suitable in the OWR setting as it enables the use of a threshold on the distance to perform known-unknown separation. Therefore we can define our prediction function ω for target samples as:

$$\hat{y}^t = \omega(\mathbf{z}^t) = \begin{cases} \arg \min_{y_s} (d_a(\mathbf{h}_{y_s}, \mathbf{z}^t)) & \text{if } \min_{y_s} (d_a(\mathbf{h}_{y_s}, \mathbf{z}^t)) < \tau \\ \text{unknown} & \text{otherwise} \end{cases} \quad (3)$$

The value of the threshold τ is one of the most important choices for an open world approach (see Table I in [13]). In our case, we can take advantage of the feature space statistics to define the threshold. Inspired by [37], we set

$$\tau = [\text{sigmoid}(\lambda - b) + 1] \cdot \phi \cdot (\ln(\lambda) + 1) \quad (4)$$

Differently from [37] we have the constraint imposed through Eq. (2), which makes the result of the logarithm always positive. We also propose a parametric formulation for the first term (between squared brackets) thanks to which we obtain a good known-unknown balancing in all the tested benchmarks (see Fig. 3).

IV. EXPERIMENTAL SETUP

A. Datasets and experimental protocol

To assess the performance of COW, we rely on the benchmark proposed in [13], by also extending it in order to consider more recent literature and new data collections. In particular we focus on five datasets perfectly suited for our

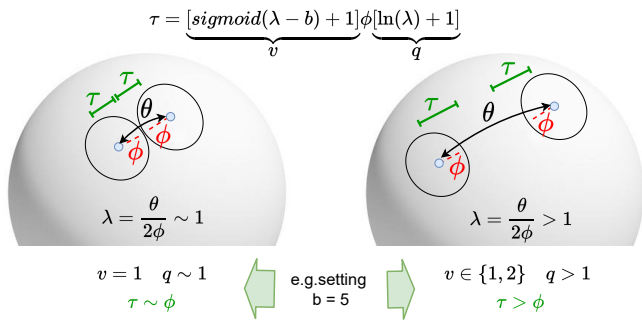


Fig. 3. Visualization of two data clusters on the learned hyperspherical embedding and of the corresponding threshold value. Tuning b means tuning v in $\{1, 2\}$. See sec. V-D for a detailed discussion.

scope. All of them contain daily-life objects (spanning from fruit and vegetables to tools and containers) recorded under very different acquisition conditions.

RGB-D Object dataset (ROD) [53] is one of the most used RGB-D dataset in robotics for object categorization. The objects were placed on a table and captured with different viewpoints. The recording was done in a strictly controlled environment without any source of noise *i.e.* without clutter, with a fixed illumination and background.

Synthetic ROD (synROD) [15] is the synthetic version of ROD, proposed to analyze the synthetic-to-real domain shift problem in a robotic context. It was recorded using publicly available 3D models rendered using a ray-tracing engine in Blender to simulate photorealistic lighting.

Autonomous Robot Indoor Dataset (ARID) [54] is a challenging dataset in which the objects were captured in a cluttered environment: the same object appears with several backgrounds, scales, views, lighting conditions, and different levels of occlusions. The purpose of this dataset was to evaluate the robustness of a recognition model when dealing with difficult but realistic scenarios.

Continual Open Set Domain Adaptation for Home Robot (COSDA-HR) [1] is a dataset composed of a source domain with hand-held objects placed in front of a uniform background and a target domain with objects captured in various natural locations in a home environment.

Continuos Object Recognition 50 (CORe50) is a collection of photos of domestic objects, captured while being held by the operator in 11 distinct sessions (8 indoor and 3 outdoor).

For what concern ROD, synROD and ARID our experimental protocol follows the same configuration proposed in [13]: among the 51 object categories that they share, we randomly consider 26 of them as known and 25 as unknown; we start with 11 known categories that increase by 5 at each incremental step for a total of four sequential tasks. For COSDA-HR instead, we follow [1]: the dataset is composed of 160 categories incrementally learned 10 at a time for a total of 16 tasks. The dataset includes a single unknown category made by a heterogeneous set of objects. CORe50 was designed to perform instance classification on 50 objects. We consider 10 of them in the first learning episode and add 5 in each of the subsequent three, keeping the last 25 as

unknown. We consider the indoor (train data) \rightarrow outdoor (test data) domain shift. To better assess the performances of the methods, for all the experiments, we consider five different random class orders and we report the obtained average.

Metrics. For the evaluation we use the same metrics used in [13]. *Acc* (Accuracy) measures the ability of the model to correctly predict the categories of the known target samples. *Acc-WR* (Accuracy Without Rejection) is similar to *Acc*, but the accuracy is computed without rejecting the target samples identified as unknown. *OWR-H* (Open World Harmonic Mean) evaluates the performance of the model as a whole, it is the harmonic mean between *Acc-WR* and the model’s accuracy in unknown sample detection.

B. Competitors

We follow [13] comparing COW against state-of-the-art methods in OWR, enhanced with single-source DG approaches to deal with the domain-shift. As competitors for the OWR setting we consider: **NNO** [14] a non-parametric approach that exploits the Nearest-Class Mean (NCM) algorithm [55] to compute the class centroids with the features extracted from a pretrained deep architecture; its more advanced version **DeepNNO** [8], in which the feature extractor is end-to-end trained and the rejection threshold is not fixed but updated during training, and **B-DOC** [9] which includes two clustering constraints in the optimization process and proposes a class-specific rejection threshold. We also include the state-of-the-art CIL method **SS-IL** [6] that uses separate softmax output layers combined with task-wise knowledge distillation to mitigate the bias toward the new classes. For both OWR and CIL the training happens in sequence on multiple tasks. The main difference is that, in CIL, the evaluation step is done on a test set composed only of images from categories learned up to the current task, in OWR the test set contains also samples from categories not learned (yet): at test time the model has to reject the unknown samples while assigning a class label only to those belonging to the learned categories. As a consequence, in CIL there is no need for the known/unknown separation threshold. To adapt SS-IL to the OWR scenario we exploit the Maximum Softmax Probability [56] of the predictions vector and to choose the threshold we rely on the logic proposed in [57].

We follow [13] also for the DG literature: a data augmentation based technique **RSDA** [17], a self-supervised based technique **RR** [15] and a method based on a regularization strategy **SC** [16]. Moreover, we include a very recent DG approach **SagNet** [18] that disentangles the sample content and style to let the network focus more on the first than on the second.

C. Implementation Details

We implement COW considering the same protocol adopted for all our competitors: we use a ResNet18 backbone trained from scratch using images of size 64×64 . When learning a new task we keep a fixed-size memory buffer to store $M = 2000$ samples of the classes of previous tasks by choosing them randomly. We train each task until our stop

TABLE II
RESULTS (%) AVERAGED OVER FIVE RANDOM CLASS ORDERS.

OWR/CIL	DG	ROD \rightarrow ARID			synROD \rightarrow ARID			synROD \rightarrow ROD			COSDA-HR			CORe50		
		Acc-WR	Acc	OWR-H	Acc-WR	Acc	OWR-H	Acc-WR	Acc	OWR-H	Acc-WR	Acc	OWR-H	Acc-WR	Acc	OWR-H
NNO [14]		18.4	3.1	5.9	16.2	7.8	13.7	21.3	13.3	21.1	8.2	3.7	7.0	15.0	0.6	1.2
DeepNNO [8]		21.3	7.3	13.4	15.9	5.4	10.0	24.4	9.6	17.0	15.1	8.2	12.8	17.0	4.5	8.1
B-DOC [9]		22.3	10.0	17.5	16.5	2.2	4.3	27.6	5.2	9.9	13.2	0.7	1.3	15.7	2.2	3.8
SS-IL [6]		17.6	14.7	16.4	21.3	9.6	16.1	29.3	16.9	22.9	8.4	5.0	6.2	23.2	16.8	18.6
NNO [14]	+ RR [15]	27.1	13.6	21.7	15.8	7.2	12.5	25.9	17.1	23.8	8.2	3.2	6.5	14.5	0.4	0.9
DeepNNO [8]		33.5	16.0	25.8	14.2	4.9	9.3	34.1	15.4	25.2	13.8	6.9	11.0	17.5	5.0	8.9
B-DOC [9]		32.2	11.7	20.4	15.7	2.2	4.3	35.9	9.7	17.3	12.3	0.5	1.0	19.5	4.5	7.1
SS-IL [6]		17.3	13.0	17.9	19.7	6.6	11.6	30.9	16.2	23.7	7.8	1.4	2.6	20.2	9.7	12.8
NNO [14]	+ SC [16]	14.1	9.8	15.5	16.0	11.6	16.9	21.9	18.8	21.2	6.4	4.6	7.2	13.3	2.9	5.3
DeepNNO [8]		20.9	15.9	22.0	15.5	8.4	14.6	25.9	17.0	25.3	15.3	11.7	15.5	18.2	8.8	13.6
B-DOC [9]		19.6	13.1	20.4	16.5	10.0	16.1	26.7	18.0	23.2	13.0	1.9	3.3	17.1	4.7	6.8
SS-IL [6]		15.2	12.9	14.7	19.0	7.9	13.2	26.8	14.6	21.0	9.0	5.5	6.6	19.3	14.7	16.0
NNO [14]	+ RSDA [17]	25.0	12.8	20.7	16.3	8.6	14.4	26.7	18.4	24.5	8.9	2.1	3.9	22.1	13.1	16.1
DeepNNO [8]		33.3	14.9	24.6	15.3	4.2	8.0	34.2	14.0	23.5	18.4	10.9	17.1	38.0	20.8	30.7
B-DOC [9]		31.9	12.2	21.1	16.3	2.5	4.9	37.9	10.8	19.1	18.2	0.6	1.0	41.4	9.8	15.9
SS-IL [6]		29.9	24.4	24.1	20.3	7.6	12.8	38.7	23.6	30.3	17.8	8.3	12.6	38.1	25.9	30.6
NNO [14]	+ SagNet [18]	19.1	3.9	7.4	15.2	7.3	12.7	20.3	12.4	19.4	8.1	3.2	6.4	15.9	2.5	4.7
DeepNNO [8]		22.5	8.7	15.5	13.7	4.7	8.8	17.9	7.1	12.8	8.5	3.9	7.2	19.3	7.6	12.4
B-DOC [9]		23.7	10.7	18.2	18.2	4.6	8.5	28.9	9.1	16.1	11.3	0.5	1.1	17.3	3.6	5.9
SS-IL [6]		24.9	19.4	21.8	20.9	8.9	14.9	32.8	17.7	24.5	8.3	3.7	6.1	27.3	17.6	23.5
COW		34.0	18.8	28.6	29.8	21.3	28.1	34.1	24.0	30.7	20.1	16.2	21.4	33.9	23.9	32.9

training condition (Eq. 2) is matched. COW has only two hyperparameters, ϵ and b , and is robust to their value as discussed in Sec. V-D.

V. EXPERIMENTS

In this section, we report the results obtained evaluating COW against the competitors introduced before. We consider different levels of domain-shift between source and target domain (we use the notation *source* \rightarrow *target*). We show how existing OWR and CIL solutions are far from solving the task of Cross-Domain Open World Recognition even if enhanced with DG approaches to bridge the domain gap. Instead, COW with a single loss and without any additional module to mitigate the domain shift outperforms the current state-of-the-art.

A. Results

In the upper part of Table II we report the results obtained using vanilla state-of-the-art OWR and CIL approaches to solve the cross-domain OWR task without the help of DG methods. Fontanel *et al.* [13] already showed how these methods perform poorly in a cross-domain scenario. The second block of the table presents the results obtained combining the OWR and CIL approaches with single-source DG methods. We re-ran¹ the experiments originally presented in [13], also considering the most recent methods SS-IL and SagNet. In the last row of the table, we show the results obtained by COW, the only approach that handles all the challenges of the CD-OWR setting at once without the need to integrate extra adaptive modules.

We now discuss the results referring to the OWR-H metric that was shown to be the most appropriate one to evaluate an open-set approach [58], [9]. For what concerns the ROD \rightarrow ARID and synROD \rightarrow ROD domain shifts, as well as

¹Our results are consistent with those of [13], made exception for few cases. An issue affected some of the numbers: we identified and corrected the problem via private communications with the authors.

the CORe50 dataset, we can generally see an improvement after adding each one of the DG approaches. There are only a few exceptions like SS-IL+SC whose combination may be incompatible: they both exploit the value of the gradient to formulate their solutions, but from two different points of view that might disagree. Anyway, the mean improvement gained with the addition of the DG strategies confirms the generalization failure of the existing OWR/CIL approaches that need auxiliary loss functions in order to properly work on a target domain different from the training one. We also observe that among the considered DG strategies the one that most often produces the highest results is the data augmentation-based approach RSDA. This confirms the great advantage that a strong data augmentation can provide in knowledge generalization [17], [41]. However in some edge cases, it may be not enough: in the synROD \rightarrow ARID shift, and in COSDA-HR dataset, all the considered DG strategies, including RSDA, do not seem to provide a significant and consistent improvement. In these cases, the domain shift is quite severe since it includes a (realistic) target domain whose images have been recorded in a cluttered environment, which is very different from the neat (and possibly synthetic) training set. Indeed the considered DG strategies are not suited to reduce such a large domain gap. As Table II clearly shows, COW obtains the best results over all the experiments, proving its effectiveness.

B. Domain Generalization through Contrastive Learning

As described in Sec. II-C, contrastive learning relies on data augmentation and the techniques usually adopted to create data variants are the same as RSDA, plus random resized crop (RC). To understand whether the advantage of COW originates mainly from this augmentation, or from the specific way in which it is used by the contrastive loss, we decided to provide other baselines with the additional RC augmentation procedure. In Table III we consider the synROD \rightarrow ARID domain shift. We compare against our

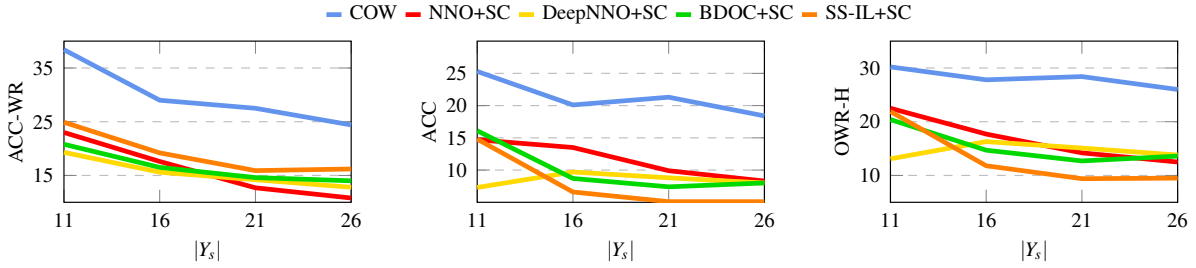


Fig. 4. Performance analysis at subsequent learning episodes for synROD \rightarrow ARID. The number of known classes $|Y_s|$ increases and the plots show how COW maintains a consistent gain over all the competitors.

TABLE III
CONTRASTIVE LEARNING VS DATA AUGMENTATION.

OWR/CIL	DG	synROD \rightarrow ARID		
		Acc-WR	Acc	OWR-H
NNO [14]	+ SC [16] + RC	16.5	13.8	14.6
NNO [14]		17.3	12.5	14.5
DeepNNO [8]	+ RSDA [17] + RC	20.4	10.7	17.7
B-DOC [9]		23.8	13.3	20.1
SS-IL [6]		27.0	11.2	17.9
COW		29.8	21.3	28.1

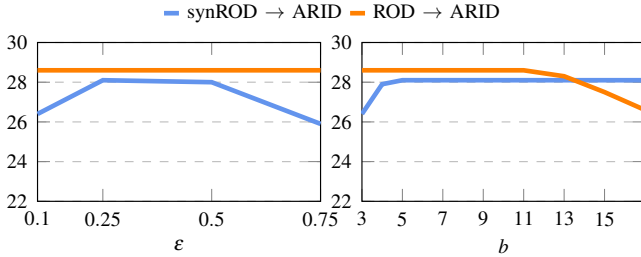


Fig. 5. OWR-H results when varying ϵ and b .

best competitor (for this shift) NNO+SC, and the methods that already use a strong augmentation because integrated with RSDA. We can observe that the performance of the considered competitors does not always increase with this addition: clearly, this procedure can also be detrimental to methods that are not designed to manage it. Nevertheless, even when the performance of the competitors increases, COW still keeps the best results. This evidences that the data augmentation is not enough and the contrastive logic is a fundamental component to enable generalization.

C. Incremental learning performance

The performance over subsequent incremental steps is an important aspect to consider when comparing incremental learning methods. We perform this analysis by reporting the scores for all the three metrics (ACC-WR, ACC, OWR-H) in Figure 4 on the synROD \rightarrow ARID domain shift. In this case, by looking at Table II, we identify the Self Challenging (SC) approach as the DG method providing the higher mean improvement to COW’s competitors, therefore we select it for the comparison. For all the three metrics we can see that COW keeps a large gap of performance over the others for all the incremental steps. Despite the natural decrease in accuracy after the first task, mainly in the ACC-WR metric, in the subsequent tasks COW is able to maintain a quite stable ACC and OWR-H performance showing a great ability

to balance the accuracy on known and unknown samples. We remark that COW exploits a very simple replay strategy with a fixed-size buffer containing randomly chosen samples of old tasks, thus the results are not the consequence of a sophisticated incremental technique.

D. Sensitivity Analysis

In Fig. 5 we evaluate the robustness of COW when changing the values for its two hyperparameters ϵ and b . They have different influence on the model performance, contingent on the value of λ , which in turns depends on the statistics of the training dataset. We consider two different shifts representing two possible extreme cases: for synROD \rightarrow ARID we have $\lambda \approx 1 + \epsilon$ (a situation similar to Fig. 3 left), while for ROD \rightarrow ARID we have $\lambda \gg 1$ (Fig. 3 right). The value of ϵ controls the minimum margin between known class clusters imposed through the stop training criterion of Eq. (2). A larger ϵ will push the training towards a larger margin by increasing clusters compactness and separation. While this is desirable, a too high value may lead to overfitting and exceptionally long times of training in order to meet the stop training condition. The value of ϵ does not influence the performance on ROD \rightarrow ARID as for this shift the margin is naturally quite high. The second hyperparameter b tunes our known-unknown separation threshold τ (see Eq. (4)). It allows to find a good balance between known and unknown accuracy for both cases. It can be noticed from the plots that the results obtained by COW are stable and high ($\geq 26\%$) for a reasonable range of values, always outperforming the best competitor (e.g. for synROD \rightarrow ARID, NNO + SC obtains 16.9, for ROD \rightarrow ARID, DeepNNO + RR obtains 25.8).

VI. CONCLUSIONS

In this work, we proposed the first approach able to tackle all the challenges of the CD-OWR setting at once. We demonstrated how a simple contrastive learning-based approach represents a very powerful solution for the task. We further introduced self-paced strategies to define both an appropriate stopping criterion and a good threshold for known-unknown separation, obtaining a model able to reach state-of-the-art results. Considering the importance of this topic we believe that our work can pave the way for future investigations.

Acknowledgements. Computational resources were provided by IIT (HPC infrastructure) and CINECA through

the IsC94 Tr-OSDG award under the ISCRA initiative. We also acknowledge the support of the European H2020 Elise project (www.elise-ai.eu). We thank Dario Fontanel for the useful discussions.

REFERENCES

- [1] I. Kishida, H. Chen, M. Baba, J. Jin, A. Amma, and H. Nakayama, "Object recognition with continual open set domain adaptation for home robot," in *WACV*, 2021.
- [2] A. Bendale and T. Boulton, "Towards open set deep networks," in *CVPR*, 2016.
- [3] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, "Generative openmax for multi-class open set classification," in *BMVC*, 2017.
- [4] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: incremental classifier and representation learning," in *CVPR*, 2017.
- [5] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE T-PAMI*, vol. 40, no. 12, pp. 2935–2947, 2017.
- [6] H. Ahn, J. Kwak, S. Lim, H. Bang, H. Kim, and T. Moon, "Ss-il: Separated softmax for incremental learning," in *ICCV*, 2021.
- [7] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, "Metric Learning for Large Scale Image Classification: Generalizing to New Classes at Near-Zero Cost," in *ECCV*, 2012.
- [8] M. Mancini, H. Karaoguz, E. Ricci, P. Jensfelt, and B. Caputo, "Knowledge is never enough: Towards web aided deep open world recognition," in *ICRA*, 2019.
- [9] D. Fontanel, F. Cermelli, M. Mancini, S. R. Bulò, E. Ricci, and B. Caputo, "Boosting deep open world recognition by clustering," *IEEE RAL*, vol. 5, no. 4, pp. 5985–5992, 2020.
- [10] G. Csurka *et al.*, *Domain adaptation in computer vision applications*. Springer, 2017.
- [11] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.
- [12] Y. Shu, Z. Cao, C. Wang, J. Wang, and M. Long, "Open domain generalization with domain-augmented meta-learning," in *CVPR*, 2021.
- [13] D. Fontanel, F. Cermelli, M. Mancini, and B. Caputo, "On the challenges of open world recognition under shifting visual domains," *IEEE RAL*, vol. 6, no. 2, pp. 604–611, 2021.
- [14] A. Bendale and T. Boulton, "Towards open world recognition," in *CVPR*, 2015.
- [15] M. R. Loghmani, L. Robbiano, M. Planamente, K. Park, B. Caputo, and M. Vincze, "Unsupervised domain adaptation through inter-modal rotation for rgb-d object recognition," *IEEE RAL*, vol. 5, no. 4, pp. 6631–6638, 2020.
- [16] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *ECCV*, 2020.
- [17] R. Volpi and V. Murino, "Addressing model vulnerability to distributional shifts over image transformation sets," in *ICCV*, 2019.
- [18] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, "Reducing domain gap by reducing style bias," in *CVPR*, 2021.
- [19] A. Prabhu, P. Torr, and P. Dokania, "Gdumb: A simple approach that questions our progress in continual learning," in *ECCV*, 2020.
- [20] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, *et al.*, "Overcoming catastrophic forgetting in neural networks," *PNAS*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [21] S. Valipour, C. Perez, and M. Jagersand, "Incremental learning for robot perception through hri," in *IROS*, 2017.
- [22] M. O. Turkoglu, F. B. Ter Haar, and N. van der Stap, "Incremental learning-based adaptive object recognition for mobile robots," in *IROS*, 2018.
- [23] A. Ayub and A. R. Wagner, "Tell me what this is: Few-shot incremental object learning by a robot," in *IROS*, 2020.
- [24] V. Lomonaco and D. Maltoni, "Core50: a new dataset and benchmark for continuous object recognition," in *PMLR*, 2017.
- [25] A. Ayub and A. R. Wagner, "F-siol-310: A robotic dataset and benchmark for few-shot incremental object learning," *ICRA*, 2021.
- [26] B. J. Meyer and T. Drummond, "The importance of metric learning for robotic vision: Open set recognition and active learning," in *ICRA*, 2019.
- [27] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis, "Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks," in *CVPR*, 2019.
- [28] R. Jeong, Y. Aytar, D. Khosid, Y. Zhou, J. Kay, T. Lampe, K. Bousmalis, and F. Nori, "Self-supervised sim-to-real adaptation for visual robotic manipulation," in *ICRA*, 2020.
- [29] G. Angeletti, B. Caputo, and T. Tommasi, "Adaptive deep learning through visual domain localization," in *ICRA*, 2018.
- [30] J. Zhang, L. Tai, P. Yun, Y. Xiong, M. Liu, J. Boedecker, and W. Burgard, "Vr-goggles for robots: Real-to-sim domain adaptation for visual control," *IEEE RAL*, vol. 4, no. 2, pp. 1148–55, 2019.
- [31] A. D’Innocente, F. C. Borlino, S. Bucci, B. Caputo, and T. Tommasi, "One-shot unsupervised cross-domain detection," in *ECCV*, 2020.
- [32] M. Wulfmeier, A. Bewley, and I. Posner, "Incremental adversarial domain adaptation for continually changing environments," in *ICRA*, 2018.
- [33] M. Mancini, H. Karaoguz, E. Ricci, P. Jensfelt, and B. Caputo, "Kitting in the wild through online domain adaptation," in *IROS*, 2018.
- [34] H. Liu, Z. Cao, M. Long, J. Wang, and Q. Yang, "Separate to adapt: Open set domain adaptation via progressive separation," in *CVPR*, 2019.
- [35] Q. Feng, G. Kang, H. Fan, and Y. Yang, "Attract or distract: Exploit the margin of open set," in *ICCV*, 2019.
- [36] Y. Pan, T. Yao, Y. Li, C.-W. Ngo, and T. Mei, "Exploring category-agnostic clusters for open-set domain adaptation," in *CVPR*, 2020.
- [37] S. Bucci, F. Cappio Borlino, B. Caputo, and T. Tommasi, "Distance-based hyperspherical classification for multi-source open-set domain adaptation," in *WACV*, 2022.
- [38] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *ICML*, 2015.
- [39] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *ICCV*, 2017.
- [40] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, no. 1, pp. 2096–2030, 2016.
- [41] F. C. Borlino, A. D’Innocente, and T. Tommasi, "Rethinking domain generalization baselines," in *ICPR*, 2021.
- [42] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *AAAI*, 2018.
- [43] S. Bucci, A. D’Innocente, Y. Liao, F. M. Carlucci, B. Caputo, and T. Tommasi, "Self-supervised learning across domains," *IEEE T-PAMI*, 2021.
- [44] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.
- [45] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.
- [46] F. Wang and H. Liu, "Understanding the behaviour of contrastive loss," in *CVPR*, 2021.
- [47] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2021.
- [48] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *NeurIPS*, 2020.
- [49] J. Tack, S. Mo, J. Jeong, and J. Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," in *NeurIPS*, 2020.
- [50] Y. Zhang, B. Hooi, D. Hu, J. Liang, and J. Feng, "Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning," in *NeurIPS*, 2021.
- [51] Z. Mai, R. Li, H. Kim, and S. Sanner, "Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning," in *CVPRW*, 2021.
- [52] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *ICML*, 2020.
- [53] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *ICRA*, 2011.
- [54] M. R. Loghmani, B. Caputo, and M. Vincze, "Recognizing objects in-the-wild: Where do we stand?" in *ICRA*, 2018.
- [55] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, "Metric learning for large scale image classification: Generalizing to new classes at near-zero cost," in *ECCV*, 2012.
- [56] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *ICLR*, 2017.

- [57] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *NeurIPS*, 2017.
- [58] S. Bucci, M. R. Loghmani, and T. Tommasi, "On the effectiveness of image rotation for open set domain adaptation," in *ECCV*, 2020.