Prototype-Guided Continual Adaptation for Class-Incremental Unsupervised Domain Adaptation

Hongbin Lin[1,3*], Yifan Zhang[2*], Zhen Qiu[1*], Shuaicheng Niu[1], Chuang Gan[4], Yanxia Liu[1†], and Mingkui Tan[1,5†]

[1]South China University of Technology [2] National University of Singapore
[3] Information Technology R&D Innovation Center of Peking University
[4] MIT-IBM Watson AI Lab
[5] Key Laboratory of Big Data and Intelligent Robot, Ministry of Education
{sehongbinlin,seqiuzhen,sensc}@mail.scut.edu.cn,
yifan.zhang@u.nus.edu, ganchuang1990@gmail.com,
{cslyx,mingkuitan}@scut.edu.cn

Abstract. This paper studies a new, practical but challenging problem, called *Class-Incremental Unsupervised Domain Adaptation* (CI-UDA), where the labeled source domain contains all classes, but the classes in the unlabeled target domain increase sequentially. This problem is challenging due to two difficulties. First, source and target label sets are inconsistent at each time step, which makes it difficult to conduct accurate domain alignment. Second, previous target classes are unavailable in the current step, resulting in the forgetting of previous knowledge. To address this problem, we propose a novel *Prototype-guided Continual Adaptation* (ProCA) method, consisting of two solution strategies. 1) Label prototype identification: we identify target label prototypes by detecting shared classes with cumulative prediction probabilities of target samples. 2) Prototype-based alignment and replay: based on the identified label prototypes, we align both domains and enforce the model to retain previous knowledge. With these two strategies, ProCA is able to adapt the source model to a class-incremental unlabeled target domain effectively. Extensive experiments demonstrate the effectiveness and superiority of ProCA in resolving CI-UDA. The source code is available at `https://github.com/Hongbin98/ProCA.git`.

Keywords: Domain Adaptation; Class-incremental Learning

1 Introduction

Unsupervised domain adaptation (UDA) seeks to improve the performance on an unlabeled target domain by leveraging a label-rich source domain via knowledge transfer [5,6,8,12,15,31,47,55]. The key challenge of UDA is the distributional

* Authors contributed equally.
† Corresponding authors.

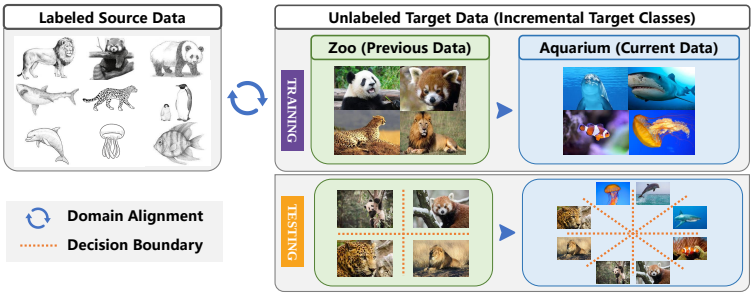

Fig. 1. An illustration of Class-incremental Unsupervised Domain Adaptation (CI-UDA), where labeled source data are accessible all the time, while unlabeled target data come online class-incrementally. When new target classes arrive, CI-UDA seeks to align the new target data to the source domain and retain the knowledge of previous target data since previous target data are unavailable.

shift between the source and target domains [32,38,51,54,56]. To deal with this, existing UDA methods conduct domain alignment either by domain-invariant feature learning [8,52] or by image transformation [12,37].

Most existing UDA methods assume the availability of all target data in advance. However, in practice, target data often come in a streaming manner with different categories [19,28,46]. For example, a practical scenario is to transfer the knowledge of sketch images for real-world animal recognition as shown in Fig. 1, where plenty of labeled sketches are easily collected in advanced but unlabeled real-world images come incrementally (*e.g.,* the images of the land animals in the zoo come first, followed by the sea animals in the aquarium). In such a scenario, it is more appropriate to adapt the source model with the target images observed so far (from partial animal classes) instead of waiting for the images of all animals to be available, which can be more memory-efficient and time-efficient. That is, the model needs to be first adapted with the target images from land animals, and then with the images from sea animals. Note that when adapting the source model with sea animals, the previous samples of land animals are unavailable for saving the data storage cost. In this scenario, existing UDA methods that assume all target classes to be available in advance tend to fail. To address this, we explore a new and practical task, called *Class-Incremental Unsupervised Domain Adaptation* (CI-UDA), where the labeled source samples are available all the time, but the unlabeled target samples come incrementally and only partial target classes are available at a time.

CI-UDA has two characteristics: 1) the target categories at the current time step are never seen before and only occupy a subspace of the source label space; 2) the target samples of previously seen categories will be unavailable for later adaptation. As a result, besides the common challenge of domain shifts in UDA [8,40], CI-UDA poses two new challenges. The first is how to detect the shared classes between source and target domains in each time step. Since only a portion of target data is available at each time step, the label space of the target domain

is inconsistent with and is partial of the source label space at each step, which makes domain alignment difficult. The second is how to alleviate catastrophic forgetting [43] of the old-class knowledge when learning new target classes. Since previous target samples are unavailable, later adaptation with new target classes results in knowledge forgetting of previous classes.

In CI-UDA, the key is to continually conduct domain adaptation in the absence of previous target samples. To deal with knowledge forgetting, a recent work [33] has shown that storing image prototypes for previous classes helps to retain knowledge. In addition, feature prototypes can also be used for domain alignment [29]. In other words, label prototypes open an opportunity for handling all challenges, simultaneously. However, a simple combination of existing methods [29,32,33] is not feasible for CI-UDA, since obtaining image prototypes for knowledge retaining [33] requires data labels but the target domain in CI-UDA is totally unlabeled. Moreover, feature prototypes [29,32] cannot update the feature extractor, so simply detecting them is unable to overcome the knowledge forgetting issue of the feature extractor in CI-UDA.

To better handle CI-UDA, we develop a new Prototype-guided Continual Adaptation (ProCA) method. To be specific, ProCA presents two solution strategies: 1) a label prototype identification strategy: we identify target label prototypes by detecting the shared class between source and target domains. Note that identifying label prototypes is challenging due to the inconsistent class space between the source and target domains. Therefore, detecting the shared classes is important, but is unfortunately difficult due to the absence of target labels. To overcome this, we dig into the difference between the shared classes and source private classes, and empirically observe (c.f. Fig. 3) that the cumulative probabilities of the shared classes are often higher than those of source private classes. Following this finding, we exploit the cumulative probabilities of target samples to detect the shared classes, and use the detected shared classes to identify target label prototypes. 2) a prototype-based alignment and replay strategy: based on the identified label prototypes, we conduct domain adaptation by aligning each target label prototype to the source center with the same class, and overcome catastrophic forgetting by enforcing the model to retain knowledge carried by the label prototypes learned from previous categories.

Extensive experiments on three benchmark datasets (*i.e.,* Office-31-CI, Office-Home-CI and ImageNet-Caltech-CI) demonstrate that ProCA is capable of handling CI-UDA. Moreover, we empirically show that ProCA can be used to improve existing partial UDA methods for tackling CI-UDA, which verifies the applicability of our method.

We summarize the main contributions of this paper as follows:

- We study a new yet difficult problem, called Class-incremental Unsupervised Domain Adaptation (CI-UDA), where unlabeled target samples come incrementally and only partial target classes are available at a time. Compared to vanilla UDA, CI-UDA does not assume all target data to be known in advance, and thus opens the opportunity for tackling more practical UDA scenarios in the wild.

– We propose a novel ProCA to handle CI-UDA. By innovatively identifying target label prototypes, ProCA is able to alleviate both domain discrepancies via prototype-based alignment and catastrophic forgetting via prototype-based knowledge replay. Moreover, ProCA can be applied to enhance existing partial domain adaptation methods to overcome CI-UDA.

2 Related Work

We first review the literature of unsupervised domain adaptation, including closed-set unsupervised domain adaptation, partial domain adaptation and continual domain adaptation. After that, we discuss a more relevant task, *i.e.,* class-incremental domain adaptation. Due to the page limit, we provide the literature of universal domain adaptation [49] and the difference between our ProCA and existing methods [1,2,3,29,32,33,43] in the supplementary (see Appendix A).

2.1 Unsupervised Domain Adaptation

Closed-set unsupervised domain adaptation (UDA). The goal of UDA [7,21,26,27,48,52] is to improve the model performance on the unlabeled target domain based on a label-rich relevant source domain. In this field, the most common task is closed-set UDA [30] which assumes that source and target domains share the same set of classes. Existing UDA methods have shown great progress in alleviating domain shifts by matching high-order moments of distributions [4,16,41], by learning domain-invariant features in an adversarial manner [8,13,36,52], or by image transformation via generative adversarial models [12,37,44]. Recently, OP-GAN [45] combines UDA with self-supervised learning, involving a self-supervised module to enforce the image content consistency.

Partial domain adaptation (PDA). Compared to closed-set UDA, PDA [1] assumes that the target label set is a subset of the source label set instead of restricting the same label set. In general, PDA aims to transfer a deep model trained from a big labeled source domain to a small unlabeled target domain. To handle the inconsistent label space, most existing methods assign class-level [1] or instance-level [2] transferability weights for source samples. To reduce negative transfer caused by source private classes, BA^{3}US [24] augments the target domain to conduct balanced adversarial alignment, while DPDAN [14] aligns the positive part of the source domain to the target domain by decomposing the source domain distribution into two parts.

Continual domain adaptation (CDA). Different from the above tasks, CDA [39] assumes that more than one unlabeled target domains come sequentially, and seeks to incrementally adapt the model to each new incoming domain without forgetting knowledge on previous domains. To this end, Dlow [9] bridges source and multiple target domains by generating a continuous flow of intermediate states, while VDFR [20] proposes to replay variational domain-agnostic features to tackle the domain shift and task shift. Recently, GRCL [39] regularizes the gradient of losses to learn discriminative features and preserve the previous knowledge, respectively.

Overall, the above methods are inapplicable in CI-UDA due to two aspects. On the one hand, closed-set UDA and PDA methods rely on the assumption that all target data are available in advance. In other words, these methods take no consideration of retaining previous knowledge. On the other hand, CDA assumes that the label set of each target domain is the same as the source label set, ignoring domain-shared classes detection. As a result, they tend to fail in handling the challenging CI-UDA.

2.2 Class-incremental Domain Adaptation

Class-incremental Domain Adaptation is related to class-incremental learning (CIL) that learns a model continuously from a data stream, where the classes increase gradually and only new classes are available at each time. CIL requires the model to classify the samples of all classes observed so far. To overcome the issue of catastrophic forgetting, existing CIL methods retain the knowledge of previous classes either by storing or generating data from previous classes [3,33,43], or by preserving the relevant model weights of the previous classes [18,25,50].

Recently, researchers extend CIL to domain adaptation and study a new task, called class-incremental domain adaptation [19,46]. Specifically, this task seeks to alleviate the domain shift between domains and incrementally learn the private classes in the target domain. To this end, with *partial labeled target private samples*, CIDA [19] generates class-specific prototypes and learns a target-specific latent space to obtain centroids under the source-free domain adaptation scenario, and CBSC [46] utilizes supervised contrastive learning for novel class adaptation and domain-invariant feature extraction.

The above class-incremental domain adaptation is different from CI-UDA in two aspects. 1) Goal: class-incremental domain adaptation seeks to handle the issue of learning new target private classes incrementally, while CI-UDA seeks to handle the issue of domain adaptation with a class-incremental target domain that has no target private class. 2) Target Labels: class-incremental domain adaptation requires one-shot or few-shot labeled target samples as a prerequisite, while CI-UDA assumes a totally unlabeled target domain. Thus, directly applying existing methods to solve CI-UDA is unfeasible. In contrast, ProCA conducts unsupervised domain alignment and knowledge replay by identifying target label prototypes, thus providing the first feasible solution to CI-UDA.

3 Problem Definition

Notation. Let $\mathcal{D}_s=\{(\mathbf{x}_j^s, y_j^s)|\ y_j^s \in \mathcal{C}_s\}_{j=1}^{n_s}$ denotes the source domain, where n_s is the number of source data pairs $(\mathbf{x}^s, y^s)$ and $\mathcal{C}_s$ denotes the source label set with the class number $|\mathcal{C}_s| = K$. Moreover, we denote the unlabeled target domain as $\mathcal{D}_t=\{\mathbf{x}_i\}_{i=1}^{n_t}$ with n_t target samples. $\mathcal{C}_t$ denotes the target label set.

Class-incremental unsupervised domain adaptation. Unsupervised domain adaptation (UDA) aims to transfer knowledge from a label-rich source domain $\mathcal{D}_s$ to an unlabeled target domain $\mathcal{D}_t$. The key to resolving UDA is to

conduct domain alignment for mitigating domain shift. Existing UDA methods generally assume that all target samples are accessible in advance and have a fixed label space that is the same to the source domain (*i.e.,* $\mathcal{C}_t = \mathcal{C}_s$). However, in real-world applications, target samples often come in a streaming manner, and meanwhile, the number of target categories may increase sequentially. To address this, we seek to explore a more practical task, namely Class-Incremental Unsupervised Domain Adaptation (CI-UDA), where labeled source samples are available all the time, but unlabeled target samples come incrementally and only partial target classes are available at a time. Here, we reuse $\mathcal{D}_t$ to denote the unlabeled target domain at the current time. Note that the label set of the target data in each time step is a subset of that of the source domain, *i.e.,* $\mathcal{C}_t \subset \mathcal{C}_s$.

Besides the domain shift that all UDA methods resolve, CI-UDA poses two new challenges: 1) how to identify the shared classes between two domains in each time step; 2) how to alleviate knowledge forgetting of old classes when learning new target classes. Due to the integration of these challenges, existing UDA methods [8,12,21,31,40,48] are incapable of handling CI-UDA. Therefore, how to handle this practical yet difficult task remains an open question.

4 Prototype-guided Continual Adaptation

Previous studies have shown that label prototypes are effective in independently handling either UDA [13,29,32] or class-incremental learning [3,33,43] tasks. Although these methods cannot be directly used to handle CI-UDA, they inspire us to explore a unified prototype-based method to handle all challenges in CI-UDA, simultaneously. This idea, however, is non-trivial to explore in practice. Since the source and target domains have different label spaces at different time steps, it is difficult to identify target label prototypes. To address these challenges, we propose a novel Prototype-guided Continual Adaptation (ProCA) method.

Method overview. We summarize the overall training scheme of ProCA in Fig. 2. ProCA consists of two solution strategies, that are, 1) label prototype identification and 2) prototype-based alignment and replay. We first briefly introduce the two strategies below.

First, we develop a label prototype identification strategy (c.f. Section 4.1) to identify target label prototypes at each time step under inconsistent label spaces between source and target domains. To this end, we firstly propose a shared class detection method to distinguish the domain-shared classes from the source private classes. Based on the detected shared label set and the target pseudo labels generated by clustering, we identify target label prototypes for each shared class and construct an adaptive memory bank $\mathcal{P}$ to record them.

Second, based on the identified label prototypes, we propose a prototype-based alignment and replay strategy (c.f. Section 4.2) to align each image prototype to the corresponding source center and enforce the model to retain knowledge learned on previous classes. Specifically, we conduct prototype-based alignment by training the feature extractor G to learn domain-invariant features through a supervised contrastive loss $\mathcal{L}_{con}$. Meanwhile, we impose a knowledge

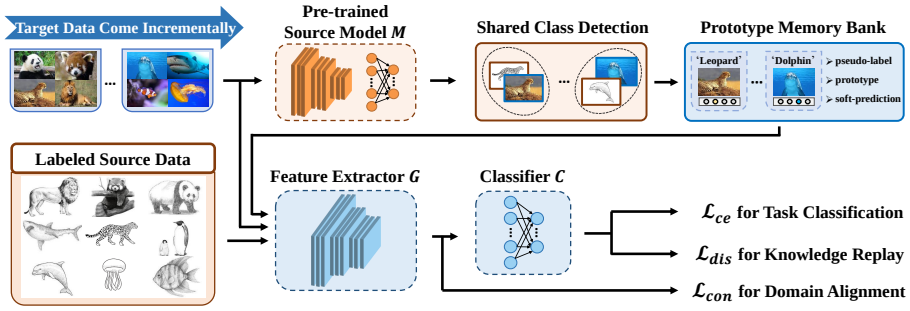

Fig. 2. An overview of Prototype-guided Continual Adaptation, ProCA consists of two strategies: 1) Label prototype identification: by detecting the shared classes between source and target domains at each time step, we identify target label prototypes for each class and record them in a memory bank. 2) Prototype-based alignment and replay: we align each target label prototype to the corresponding source center for training a domain-invariant feature extractor via $\mathcal{L}_{con}$; meanwhile, we use the saved label prototypes to enforce the model to retain knowledge learned from previous classes via $\mathcal{L}_{dis}$. Moreover, we use the cross-entropy $\mathcal{L}_{ce}$ for task classification based on the labeled source samples and pseudo-labeled target samples.

distillation loss $\mathcal{L}_{dis}$ for prototype-based knowledge replay. Moreover, based on the pseudo-labeled target data and labeled source data, we train the whole model $\{G, C\}$ via the standard cross-entropy loss $\mathcal{L}_{ce}$.

Overall, the training objective of ProCA is as follows:

$$\min_{\{\theta_g,\theta_c\}} \mathcal{L}_{ce}(\theta_g,\theta_c) + \lambda\mathcal{L}_{con}(\theta_g) + \eta\mathcal{L}_{dis}(\theta_g,\theta_c), \tag{1}$$

where θ_g and θ_c denote the parameters of the feature extractor G and the classifier C, respectively. Moreover, λ and η are trade-off parameters.

4.1 Label Prototype Identification

The key step in our proposed ProCA is to identify target label prototypes, which is non-trivial in the setting of CI-UDA. To this end, we propose a label prototype identification strategy that consists of four components: 1) shared class detection; 2) pseudo label generation for target data; 3) prototype memory bank construction and 4) prototype memory bank updating.

Shared class detection. When new unlabeled target samples arrive, it is difficult to detect the shared classes between the source and target domains since the target samples are unlabeled. To resolve this, we dig into the difference of the pre-trained source model in predicting the shared classes and the source private classes. As shown in Fig. 3, we find that the cumulative prediction probabilities of the target samples regarding the shared classes are higher than those regarding source private classes. Following this, we propose to detect the shared classes based on the cumulative probabilities of target samples. Specifically, as shown

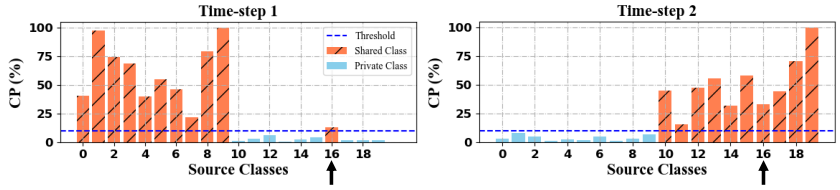

Fig. 3. The cumulative probability (CP) of target samples regarding source classes on Art→Real World, Office-Home-CI. For each time step, 10 target classes are available, *i.e.,* Class 0 to 9 in time step 1 and Class 10 to 19 in time step 2. The results show that the CP values of the shared classes are often higher than those of source private classes. Moreover, during training, since the CP of the 16-th class in time step 2 is higher than that in time step 1, we update the corresponding target class prototypes.

in Fig. 4, we exploit the source pre-trained model M to infer all target samples in each time step and obtain the cumulative probability of each class k by:

$$u_k = \sum_{i=1}^{n_t} C_k(G(\mathbf{x}_i)), \tag{2}$$

where $C_k(\cdot)$ denotes the k-th element in the softmax output prediction and n_t denotes the number of target samples at the current time. To enhance the generalization, we normalize the cumulative probability u_k to $[0, 1]$ by $u_k{=}\frac{u_k - min(\mathbf{u})}{max(\mathbf{u}) - min(\mathbf{u})}$, where $\mathbf{u}{=}[u_0, u_1, ..., u_K]$ is the probability vector in terms of all K classes.

Based on the cumulative probability u_k and a pre-defined threshold α, the judgement of the class k is made by: if $u_k \geq \alpha$, class k is a shared class; otherwise, class k is a source private class.

Pseudo label generation for target data. Based on the identified shared classes, we next generate pseudo labels for unlabeled target samples with a self-supervised pseudo-labeling strategy [23]. To be specific, let $\mathbf{q}_i{=}G(\mathbf{x}_i)$ be the extracted feature w.r.t. $\mathbf{x}_i$ and let $\hat{y}_i^k{=}C_k(\mathbf{q}_i)$ be the predicted probability of the classifier regarding class k, we first attain the initial centroid for each class k in the shared label set by:

$$\mathbf{c}_k = \frac{\sum_{i=1}^{n_t} \hat{y}_i^k \mathbf{q}_i}{\sum_{i=1}^{n_t} \hat{y}_i^k}. \tag{3}$$

Such an initialization is able to characterize well the distribution of different categories [23]. Based on these centroids, the pseudo label of the i-th target data is obtained via a nearest centroid approach:

$$\bar{y}_i = \arg\max_k \phi(\mathbf{q}_i, \mathbf{c}_k), \tag{4}$$

where $\phi(\cdot, \cdot)$ denotes the cosine similarity, and the pseudo label $\bar{y}_i \in \mathbb{R}^1$ is a scalar. During pseudo label generation, we update the centroid of each class by $\mathbf{c}_k = \frac{\sum_{i=1}^{n_t} \mathbb{I}(\bar{y}_i{=}k)\mathbf{q}_i}{\sum_{i=1}^{n_t} \mathbb{I}(\bar{y}_i{=}k)}$ and then update pseudo labels based on Eqn. (4) one more

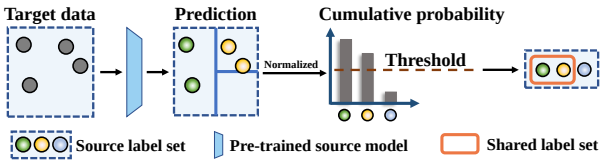

Fig. 4. The process of shared class identification. We first compute the cumulative probabilities by summing the output predictions of the pre-trained source model regarding all target samples. Then, we rescale the cumulative probabilities to $[0, 1]$ by min-max normalization. Based on the normalized probabilities, we judge the shared classes through thresholding.

time, where $\mathbb{I}(\cdot)$ is the indicator function. Note that we only compute the class centroids for the shared classes.

Prototype memory bank construction. Based on the detected shared label set and the generated target pseudo labels, we then identify target label prototypes for each shared class. Specifically, we maintain a memory bank $\mathcal{P} = \{(\mathbf{p}_i, \mathbf{h}_i, \bar{y}_i)\}_{i=1}^{N}$ to record prototypes for all detected shared classes, where $\mathbf{p}_i$, $\mathbf{h}_i$, $\bar{y}_i$ and N denote the image prototype, the predicted soft label, the predicted hard pseudo label and the number of prototypes, respectively. Moreover, we denote all seen target label set as $\mathcal{C}_{at}$ and save M image prototypes for each class in the memory bank, *i.e.,* $N = |\mathcal{C}_{at}|M$. During the training process, when a new pseudo-labeled target class comes, we expand the memory bank by adding the corresponding target prototypes. Formally, for the k-th class, we denote the pseudo-labeled target domain as $\mathcal{D}_t^k = \{\mathbf{x}_i^k\}_{i=1}^{n_k}$ and attain its feature center by $\mathbf{f}_t^k = \frac{1}{n_k}\sum_{i=1}^{n_k} G(\mathbf{x}_i^k)$. Inspired by iCaRL [33], we select the image prototype for the k-th class via a nearest neighbor approach based on the target feature center:

$$\mathbf{p}_m^k = \arg\min_{\mathbf{x}^k \in \mathcal{D}_t^k} \left\| \mathbf{f}_t^k - \frac{1}{m}[G(\mathbf{x}^k) + \sum_{i=1}^{m-1} G(\mathbf{p}_i^k)] \right\|_2, \tag{5}$$

where m is the iterative index range from 1 to M. Note that we iterate Eqn. (5) for M times to obtain M prototypes.

Prototype memory bank updating. During the shared class detection process, false shared classes may exist, disturbing pseudo label generation for target data. In this sense, the image prototypes of these classes need to be updated. To this end, we devise an updating strategy based on the cumulative probability u_k. Specifically, for a target class k existing in the memory bank, we update its prototypes when a higher u_k occurs. For example, in Fig. 3, when the cumulative probability of the 16-th class in the time step 2 is higher than that in the time step 1, the image prototypes of the 16-th class would be updated via Eqn. (5).

Algorithm 1 Training paradigm of ProCA

Require: Unlabeled target data $\mathcal{D}_t$=$\{\mathbf{x}_i\}_{i=1}^{n_t}$ at the current time; Pre-trained source model $\{G, C\}$; Prototype memory bank $\mathcal{P}$; Training epoch E; Parameters η, λ, α, T.

```
Detect shared classes based on Eqn. (2);
Obtain target pseudo labels based on Eqn. (4);
for e = 1 → E do
    Update target label prototypes based on Eqn. (5);
    Extract target data features G(x) based on G;
    Obtain target predictions C(G(x)) based on C;
    Compute L_con, L_dis based on Eqns. (7) and (8);
    Update G and C by optimizing Eqn. (1)
end for
return G and C.
```

4.2 Prototype-based Alignment and Replay

Based on target label prototypes, we develop a new prototype-based alignment and replay strategy to handle the issues of domain shifts and catastrophic forgetting below.

Prototype-based domain alignment. Based on the target label prototypes, we are able to conduct class-wise alignment to explicitly mitigate domain shifts. To this end, we propose to align each pseudo-labeled prototype to the source center of the corresponding class. To be specific, for the k-th class, we first attain source feature center $\mathbf{f}_s^k$ by:

$$\mathbf{f}_s^k = \frac{\sum_{j=1}^{n_s} \mathbb{I}(y_j^s = k) G(\mathbf{x}_j^s)}{\sum_{j=1}^{n_s} \mathbb{I}(y_j^s = k)}. \tag{6}$$

Then for any image prototype $\mathbf{p}_i$ as an anchor, we conduct prototype alignment via the contrastive loss [53,17]:

$$\mathcal{L}_{con} = -\log \frac{\exp(\mathbf{v}_i^\top \mathbf{f}_s^{\bar{y}_i}/\tau)}{\exp(\mathbf{v}_i^\top \mathbf{f}_s^{\bar{y}_i}/\tau) + \sum_{j=1, j\neq\bar{y}_i}^{K-1} \exp(\mathbf{v}_i^\top \mathbf{f}_s^{j}/\tau)}, \tag{7}$$

where $\mathbf{v}_i$=$G(\mathbf{p}_i)$ denotes the extracted feature of the image prototype $\mathbf{p}_i$ with $\bar{y}_i$ as its pseudo label and τ denotes a temperature factor. This loss enables the feature extractor G to learn domain-invariant features, which helps to alleviate domain discrepancies.

Prototype-based knowledge replay. Since target samples of previous classes are unavailable, the model suffers from catastrophic forgetting [43] during CI-UDA. To overcome this, based on the identified prototypes with soft labels, we adopt knowledge distillation [22] to enforce the model to retain the knowledge acquired from previous classes:

$$\mathcal{L}_{dis} = -\frac{1}{N} \sum_{i=1}^{N} \mathbf{h}_i^\top \log C(G(\mathbf{p}_i)), \tag{8}$$

where N denotes the number of prototypes.

At last, we summarize the pseudo-code of ProCA in Algorithm 1, while the pseudo-code of the prototype identification scheme is put in Appendix B.

5 Experiments

5.1 Experimental Setup

To verify the effectiveness of the proposed method, we conduct empirical studies based on the following experimental settings.

Datasets. We construct three dataset variants to simulate class-incremental scenarios, based on benchmark UDA datasets, *i.e.,* Office-31 [35], Office-Home [42], and ImageNet-Caltech [10,34]. 1) ***Office-31-CI*** consists of three distinct domains, *i.e.,* Amazon (A), Webcam (W) and DSLR (D). Three domains share 31 categories. We divide each domain into three disjoint subsets with each containing 10 categories in alphabetic order. 2) ***Office-Home-CI*** contains four distinct domains, *i.e.,* Artistic images (Ar), Clip Art (Cl), Product images (Pr) and Real-world images (Rw), each with 65 categories. For each domain, we build six disjoint subsets with 10 categories in random order. 3) ***ImageNet-Caltech-CI*** includes ImageNet-1K [34] and Caltech-256 [10]. Based on the shared 84 classes, we form two tasks: ImageNet (1000) $\rightarrow$ Caltech (84) and Caltech (256) $\rightarrow$ ImageNet (84). For target domains, we build eight disjoint subsets with each containing 10 categories. More details of data construction are in Appendix C.

Implementation details. We implement our method in PyTorch and report the mean $\pm$ stdev result over 3 different runs. ResNet-50 [11], pre-trained on ImageNet, is used as the network backbone. In ProCA, we train the model using the SGD optimizer with a learning rate of 0.001. In addition, the training epochs are set to 10 for Office-31-CI, 30 for Office-Home-CI, and 15 for ImageNet-Caltech-CI, respectively. For hyper-parameter, we set λ, η, α and M to 0.1, 1, 0.15 and 10. More training details of ProCA are in Appendix D.

Compared methods. We compare ProCA with four categories of baselines: (1) source-only: ResNet-50 [11]; (2) unsupervised domain adaptation: DANN [8]; (3) partial domain adaptation: PADA [1], ETN [2], BA^{3}US [24]; (4) class-increme ntal domain adaptation: CIDA [19].

Evaluation protocols. To fully evaluate the proposed method, we report three kinds of accuracy measures. 1) ***Final Accuracy***: the classification accuracy in the final time step of CI-UDA. 2) ***Step-level Accuracy***: the accuracy in each time step to evaluate the ability of sequentially learning. 3) ***Final S-1 Accuracy***: the average accuracy of step-1 classes in the final time step to evaluate the ability to handle catastrophic forgetting.

5.2 Comparisons with Previous Methods

We first compare our ProCA with previous methods in terms of Final Accuracy. The results are reported in Tables 1 and 2, which give the following observations. 1) ProCA outperforms all compared methods by a large margin in terms of the averaged Final Accuracy. To be specific, ProCA achieves the best or comparable performance on all transfer tasks (*e.g.,* Ar$\rightarrow$Cl on Office-Home-CI), which demonstrates the effectiveness of our method. 2) Compared with PDA methods, *i.e.,* PADA [1], ETN [2] and BA^{3}US [24], the superior performance of our

Table 1. Final Accuracy (%) on **Office-Home-CI**. DA and CI indicate domain adaptation and class-incremental learning.

Method	DA	CI	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
ResNet-50	✗	✗	47.6	65.2	72.7	54.7	62.8	66.1	52.4	44.7	74.0	66.2	47.4	77.4	60.9
DANN [8]	✓	✗	33.1	40.0	45.8	36.8	36.6	44.1	32.0	29.8	49.8	42.4	40.2	55.2	40.5
PADA [1]	✓	✗	24.8	41.4	55.1	18.3	35.0	36.3	25.9	26.2	53.7	46.8	31.4	50.0	37.1
ETN [2]	✓	✗	42.4	2.8	7.4	4.3	60.3	6.3	50.7	33.8	70.8	3.7	43.5	75.1	33.4
BA^{3}US [24]	✓	✗	33.7	39.7	63.2	36.6	39.1	53.7	36.5	24.9	53.4	52.2	35.9	65.9	44.6
CIDA [19]	✓	✓	32.2	45.9	49.1	36.5	48.6	46.6	51.6	33.5	59.0	64.0	38.0	65.1	47.5
ProCA (ours)	✓	✓	**51.9**$_{\pm 0.4}$	**75.2**$_{\pm 0.2}$	**86.1**$_{\pm 0.3}$	**60.8**$_{\pm 0.1}$	**69.7**$_{\pm 0.1}$	**74.7**$_{\pm 0.7}$	**60.1**$_{\pm 0.2}$	**51.0**$_{\pm 0.2}$	**84.2**$_{\pm 0.4}$	**75.8**$_{\pm 0.2}$	**51.2**$_{\pm 0.5}$	**86.4**$_{\pm 0.1}$	**68.9**$_{\pm 0.1}$

Table 2. Final Accuracy (%) on **Office-31-CI** and **ImageNet-Caltech-CI**. DA and CI indicate adaptation and class-incremental learning.

Method	DA	CI	Office-31-CI							ImageNet-Caltech-CI		
			A→D	A→W	D→A	D→W	W→A	W→D	Avg.	C→I	I→C	Avg.
ResNet-50	✗	✗	74.1	74.4	58.5	96.9	61.2	99.6	77.5	72.3	70.7	71.5
DANN [8]	✓	✗	74.9	72.5	55.7	96.6	51.4	97.7	74.8	58.8	31.4	45.1
PADA [1]	✓	✗	56.9	61.5	12.5	82.4	46.7	84.3	57.4	37.3	45.9	41.6
ETN [2]	✓	✗	21.3	82.2	61.7	94.3	**64.1**	**100.0**	70.6	1.4	3.1	2.3
BA^{3}US [24]	✓	✗	74.1	73.3	63.3	94.8	64.0	**100.0**	78.3	60.8	45.0	52.9
CIDA [19]	✓	✓	70.4	64.5	48.1	95.1	52.7	98.8	71.6	69.3	49.2	59.2
ProCA (ours)	✓	✓	**81.8**$_{\pm 0.6}$	**82.5**$_{\pm 0.4}$	**65.2**$_{\pm 0.3}$	**99.1**$_{\pm 0.1}$	**64.1**$_{\pm 0.2}$	99.6$_{\pm 0.2}$	**82.1**$_{\pm 0.3}$	**82.9**$_{\pm 0.2}$	**83.1**$_{\pm 0.3}$	**83.0**$_{\pm 0.1}$

Table 3. Step-level Accuracy (%) on **Office-31-CI** and **Office-Home-CI**. DA and CI indicate adaptation and class-incremental learning.

Method	DA	CI	Office-31-CI				Office-Home-CI						
			Step 1	Step 2	Step 3	Avg.	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Avg.
ResNet-50 [11]	✗	✗	85.7	81.8	77.5	81.6	61.2	61.7	61.2	62.0	62.3	62.4	61.8
DANN [8]	✓	✗	82.4	79.6	74.8	78.9	42.7	40.5	41.1	41.1	39.8	40.5	40.9
PADA [1]	✓	✗	87.5	69.9	57.4	71.6	63.0	49.3	40.4	37.7	37.4	37.1	44.2
ETN [2]	✓	✗	**92.0**	82.7	70.6	81.8	62.7	62.0	59.2	58.7	49.0	33.4	54.2
BA^{3}US [24]	✓	✗	90.7	**85.9**	78.3	85.0	66.6	60.4	53.6	49.1	46.0	44.6	53.4
CIDA [19]	✓	✓	85.5	79.1	71.6	78.7	57.9	53.6	51.8	50.1	49.6	47.5	51.8
ProCA (ours)	✓	✓	91.3	**85.9**	**82.1**	**86.3**	**70.2**	**70.1**	**68.2**	**68.5**	**68.7**	**68.9**	**86.0**

method shows that retaining knowledge learned from previous categories is important for handling CI-UDA. 3) Since CIDA [19] also designs a regularization term to prevent catastrophic forgetting, it performs better than PDA methods in CI-UDA. However, CIDA ignores the source private classes in CI-UDA, which may result in negative transfer, and thus cannot handle CI-UDA very well. 4) Domain adaptation methods even perform worse than ResNet-50, which implies that only conducting alignment may make the model biased towards the target categories at the current step and forget the knowledge of previous categories.

We also report the Step-level Accuracy of all methods in Tables 3. If only considering the time step 1, CI-UDA degenerates to a standard PDA problem. In this case, previous PDA methods (*i.e.,* PADA [1], ETN [2] and BA^{3}US [24]) perform well. However, when learning new target samples at a new time step, these methods suffer from severe performance degradation while our ProCA maintains a relatively stable yet promising performance. To investigate the reason, we show the accuracy drop in percentage of these step-1 classes between the time step 1 and each time step. As shown in Fig. 5, when learning new target categories,

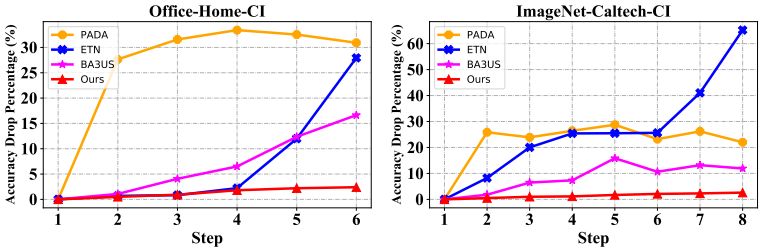

Fig. 5. Accuracy drop in percentage of different methods on **Office-Home-CI** and **ImageNet-Caltech-CI**. The accuracy drop means the accuracy difference of the step-1 classes between the time step 1 and the following each time step. The results show that our method has a significantly smaller accuracy drop in percentage than the compared methods, which shows that our method is skilled at alleviating catastrophic forgetting.

Table 4. Comparisons of the existing partial domain adaptation methods with and without our label prototype identification strategy on **Office-31-CI**. We show the final accuracy (%) and final S-1 accuracy (%) of two partial domain adaptation methods.

Method	Prototypes	Metric	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
PADA	✗	Final Acc. (%)	56.9	61.5	12.5	82.4	46.7	84.3	57.4
	✓		70.8	76.2	40.9	94.6	55.6	99.8	73.0
	✗	Final S-1 Acc. (%)	35.2	49.9	17.2	74.8	39.9	72.8	48.3
	✓		79.5	80.9	62.6	96.6	62.5	100.0	80.4
BA^{3}US	✗	Final Acc. (%)	74.1	73.3	63.3	94.8	64.0	100.0	78.3
	✓		75.4	77.4	64.2	100.0	65.5	100.0	80.4
	✗	Final S-1 Acc. (%)	89.7	89.0	76.7	100.0	77.3	99.8	88.7
	✓		92.4	90.9	78.6	100.0	77.3	100.0	89.8
ProCA (ours)	✓	Final Acc. (%)	81.6	82.6	65.5	99.1	63.9	99.8	82.1
	✓	Final S-1 Acc. (%)	96.7	94.2	74.1	100.0	80.0	100.0	90.8

the absence of target samples from previous categories causes state-of-the-art PDA methods to forget previous knowledge, leading to a severe accuracy drop of step-1 classes. In contrast, ProCA handles catastrophic forgetting effectively and shows a promising result in terms of Step-level Accuracy. Due to the page limitation, we put more detailed results of *each subtask* in the three datasets in terms of the Step-level Accuracy and the Final S-1 Accuracy in Appendix H.

5.3 Application to Enhancing Partial Domain Adaptation

In this section, we seek to determine whether ProCA can be used to enhance existing PDA methods, which cannot overcome catastrophic forgetting of previously seen categories in CI-UDA. To this end, we apply ProCA to improve classic PDA methods (*i.e.,* PADA [1] and BA^{3}US [24]) by integrating them with our label prototype identification strategy. As shown in Table 4, combining with ProCA significantly increases the performance of PDA methods, which demonstrates the applicability of our method to boost existing PDA methods for handling CI-UDA. Such an observation can also be supported by the results of applying ProCA to improve ETN [2], as shown in Appendix E.

Table 5. Ablation studies of the losses (*i.e.,* $\mathcal{L}_{ce}$, $\mathcal{L}_{dis}$ and $\mathcal{L}_{con}$) in ProCA. We show the Final Accuracy (%) and the Final S-1 Accuracy (%) on the 6 tasks of Office-31-CI.

Backbone	$\mathcal{L}_{ce}$	$\mathcal{L}_{dis}$	$\mathcal{L}_{con}$	Final Acc. (%)							Final S-1 Acc. (%)						
				A→D	A→W	D→A	D→W	W→A	W→D	Avg.	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
✓				74.1	74.4	58.5	96.9	61.2	99.6	77.5	87.8	85.3	68.5	100.0	71.4	100.0	85.5
✓	✓			76.8	78.8	59.6	99.0	62.4	99.6	79.4	90.0	90.8	68.6	100.0	71.5	100.0	86.8
✓	✓	✓		79.9	82.0	64.0	99.0	62.9	99.6	81.2	93.7	93.8	71.5	100.0	75.0	100.0	89.0
✓	✓		✓	78.7	81.5	63.2	99.0	63.5	99.0	80.8	91.3	92.4	69.7	100.0	76.7	98.0	88.0
✓	✓	✓	✓	81.6	82.6	65.5	99.1	63.9	99.8	**82.1**	96.7	94.2	74.1	100.0	80.0	100.0	**90.8**

5.4 Ablation Studies

To examine the effectiveness of the losses in ProCA, we show the quantitative results of the models optimized by different losses. As shown in Table 5, introducing $\mathcal{L}_{dis}$ or $\mathcal{L}_{con}$ enhances the model performance compared to optimizing the model with $\mathcal{L}_{ce}$ only. On the one hand, such a result verifies that prototype-based knowledge replay is able to alleviate catastrophic forgetting, resulting in promoting Final S-1 Accuracy. On the other hand, it also verifies that prototype-based domain alignment is able to mitigate domain shifts, resulting in promoting Final Accuracy. When combining the losses (*i.e.,* $\mathcal{L}_{ce}$, $\mathcal{L}_{dis}$, $\mathcal{L}_{con}$) together, we obtain the best performance.

In addition, we investigate the influences of hyper-parameters. The results in Appendix F show that ProCA is non-sensitive to λ and η, and the best performance of ProCA can be usually achieved by setting $\lambda = 0.1$ and $\eta = 1$. Moreover, we recommend setting $\alpha = 0.15$ since a high threshold helps to filter domain-shared classes out. Furthermore, we also investigate the influence of the number of prototypes and incremental classes in Appendix F, and evaluate the effectiveness of our shared class detection strategy in Appendix G.

6 Conclusions

In this paper, we have explored a practical transfer learning task, namely class-incremental unsupervised domain adaptation. To solve this challenging task, we have proposed a novel Prototype-guided Continual Adaptation (ProCA) method, which presents two solution strategies. 1) Label prototype identification: we identify target label prototypes with the help of a new shared class detection strategy. 2) Prototype-based alignment and replay: based on the identified label prototypes, we resolve the domain discrepancies and catastrophic forgetting via prototype-guided contrastive alignment and knowledge replay, respectively. Extensive experiments on three benchmark datasets , *i.e.,* Office-31-CI, Office-Home-CI and ImageNet-Caltech-CI, have demonstrated the effectiveness of ProCA in handling class-incremental unsupervised domain adaptation.

Acknowledgements. This work was partially supported by National Key R&D Program of China (No.2020AAA0106900), National Natural Science Foundation of China (NSFC) 62072190, Program for Guangdong Introducing Innovative and Enterpreneurial Teams 2017ZT07X183.

References

1. Cao, Z., Ma, L., Long, M., et al.: Partial adversarial domain adaptation. In: ECCV (2018)
2. Cao, Z., You, K., Long, M., et al.: Learning to transfer examples for partial domain adaptation. In: CVPR. pp. 2985–2994 (2019)
3. Castro, F.M., Marín-Jiménez, M.J., Guil, N., et al.: End-to-end incremental learning. In: ECCV. pp. 233–248 (2018)
4. Chen, C., Fu, Z., Chen, Z., et al.: Homm: Higher-order moment matching for unsupervised domain adaptation. In: AAAI. pp. 3422–3429 (2020)
5. Chen, S., Harandi, M., Jin, X., Yang, X.: Domain adaptation by joint distribution invariant projections. IEEE Transactions on Image Processing pp. 8264–8277 (2020)
6. Chen, Y., Li, W., Sakaridis, C., et al.: Domain adaptive faster r-cnn for object detection in the wild. In: CVPR. pp. 3339–3348 (2018)
7. Du, Z., Li, J., Su, H., Zhu, L., Lu, K.: Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In: CVPR. pp. 3937–3946 (2021)
8. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML (2015)
9. Gong, R., Li, W., Chen, Y., et al.: Dlow: Domain flow for adaptation and generalization. In: CVPR. pp. 2477–2486 (2019)
10. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset (2007)
11. He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: CVPR (2016)
12. Hoffman, J., Tzeng, E., Park, T., et al.: Cycada: Cycle-consistent adversarial domain adaptation. In: ICML (2018)
13. Hu, D., Liang, J., Hou, Q., Yan, H., Chen, Y.: Adversarial domain adaptation with prototype-based normalized output conditioner. IEEE Transactions on Image Processing (2021)
14. Hu, J., Tuo, H., Wang, C., et al.: Discriminative partial domain adversarial network. In: ECCV. pp. 632–648 (2020)
15. Inoue, N., Furuta, R., Yamasaki, T., et al.: Cross-domain weakly-supervised object detection through progressive domain adaptation. In: CVPR. pp. 5001–5009 (2018)
16. Kang, G., Jiang, L., Yang, Y., et al.: Contrastive adaptation network for unsupervised domain adaptation. In: CVPR. pp. 4893–4902 (2019)
17. Khosla, P., Teterwak, P., Wang, C., et al.: Supervised contrastive learning. In: NeurIPS (2020)
18. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017)
19. Kundu, J.N., Venkatesh, R.M., Venkat, N., et al.: Class-incremental domain adaptation. In: ECCV. pp. 53–69 (2020)
20. Lao, Q., Jiang, X., Havaei, M., et al.: Continuous domain adaptation with variational domain-agnostic feature replay. ArXiv (2020)
21. Li, C., Lee, G.H.: From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In: CVPR. pp. 1482–1491 (2021)
22. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**, 2935–2947 (2018)
23. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: ICML (2020)

24. Liang, J., Wang, Y., Hu, D., et al.: A balanced and uncertainty-aware approach for partial domain adaptation. In: ECCV. pp. 123–140 (2020)
25. Liu, X., Masana, M., Herranz, L., et al.: Rotate your networks: Better weight consolidation and less catastrophic forgetting. International Conference on Pattern Recognition pp. 2262–2268 (2018)
26. Melas-Kyriazi, L., Manrai, A.K.: Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In: CVPR. pp. 12435–12445 (2021)
27. Na, J., Jung, H., Chang, H.J., Hwang, W.: Fixbi: Bridging domain spaces for unsupervised domain adaptation. In: CVPR. pp. 1094–1103 (2021)
28. Niu, S., Wu, J., Zhang, Y., et al.: Efficient test-time model adaptation without forgetting. In: ICML (2022)
29. Pan, Y., Yao, T., Li, Y., et al.: Transferrable prototypical networks for unsupervised domain adaptation. In: CVPR (2019)
30. Panareda Busto, P., Gall, J.: Open set domain adaptation. In: ICCV. pp. 754–763 (2017)
31. Pei, Z., Cao, Z., Long, M., et al.: Multi-adversarial domain adaptation. In: AAAI (2018)
32. Qiu, Z., Zhang, Y., Lin, H., et al.: Source-free domain adaptation via avatar prototype generation and adaptation. In: IJCAI (2021)
33. Rebuffi, S.A., Kolesnikov, A., Sperl, G., et al.: icarl: Incremental classifier and representation learning. In: CVPR. pp. 5533–5542 (2017)
34. Russakovsky, O., Deng, J., Su, H., et al.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
35. Saenko, K., Kulis, B., Fritz, M., et al.: Adapting visual category models to new domains. In: ECCV (2010)
36. Saito, K., Watanabe, K., Ushiku, Y., et al.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR. pp. 3723–3732 (2018)
37. Sankaranarayanan, S., Balaji, Y., Castillo, C.D., et al.: Generate to adapt: Aligning domains using generative adversarial networks. In: CVPR (2018)
38. Tang, H., Chen, K., Jia, K.: Unsupervised domain adaptation via structurally regularized deep clustering. In: CVPR (2020)
39. Tang, S., Su, P., Chen, D., et al.: Gradient regularized contrastive learning for continual domain adaptation. In: AAAI. pp. 2–13 (2021)
40. Tzeng, E., Hoffman, J., Saenko, K., et al.: Adversarial discriminative domain adaptation. In: CVPR. pp. 2962–2971 (2017)
41. Tzeng, E., Hoffman, J., Zhang, N., et al.: Deep domain confusion: Maximizing for domain invariance. ArXiv (2014)
42. Venkateswara, H., Eusebio, J., Chakraborty, S., et al.: Deep hashing network for unsupervised domain adaptation. In: CVPR (2017)
43. Wu, Y., Chen, Y., Wang, L., et al.: Large scale incremental learning. In: CVPR. pp. 374–382 (2019)
44. Xia, H., Ding, Z.: Hgnet: Hybrid generative network for zero-shot domain adaptation. In: ECCV. pp. 55–70 (2020)
45. Xie, X., Chen, J., Li, Y., et al.: Self-supervised cyclegan for object-preserving image-to-image domain adaptation. In: ECCV. pp. 498–513 (2020)
46. Xu, M., Islam, M., Lim, C.M., et al.: Class-incremental domain adaptation with smoothing and calibration for surgical report generation. In: MICCAI. pp. 269–278 (2021)
47. Yang, J., Shi, S., Wang, Z., et al.: St3d: Self-training for unsupervised domain adaptation on 3d object detection. In: CVPR. pp. 10363–10373 (2021)

48. Yang, J., Shi, S., Wang, Z., et al.: St3d: Self-training for unsupervised domain adaptation on 3d object detection. In: CVPR. pp. 10368–10378 (2021)
49. You, K., Long, M., Cao, Z., et al.: Universal domain adaptation. In: CVPR. pp. 2720–2729 (2019)
50. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: ICML. pp. 3987–3995 (2017)
51. Zhang, Y., David, P., Gong, B.: Curriculum domain adaptation for semantic segmentation of urban scenes. In: ICCV. pp. 2039–2049 (2017)
52. Zhang, Y., Chen, H., Wei, Y., et al.: From whole slide imaging to microscopy: Deep microscopy adaptation network for histopathology cancer image classification. In: MICCAI (2019)
53. Zhang, Y., Hooi, B., Hong, L., et al.: Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning. In: NeurIPS (2021)
54. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. Arxiv (2021)
55. Zhang, Y., Wei, Y., Wu, Q., Zhao, P., Niu, S., Huang, J., Tan, M.: Collaborative unsupervised domain adaptation for medical image diagnosis. IEEE Transactions on Image Processing **29**, 7834–7844 (2020)
56. Zou, Y., Yu, Z., Kumar, B.V.K.V., et al.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV (2018)