Complementary Domain Adaptation and Generalization for Unsupervised Continual Domain Shift Learning

Wonguk Cho[1], Jinha Park[2], and Taesup Kim[1*]
[1]Graduate School of Data Science, [2]Department of Electrical and Computer Engineering
Seoul National University
{wongukcho, jhpark410, taesup.kim}@snu.ac.kr

Abstract

Continual domain shift poses a significant challenge in real-world applications, particularly in situations where labeled data is not available for new domains. The challenge of acquiring knowledge in this problem setting is referred to as unsupervised continual domain shift learning. Existing methods for domain adaptation and generalization have limitations in addressing this issue, as they focus either on adapting to a specific domain or generalizing to unseen domains, but not both. In this paper, we propose Complementary Domain Adaptation and Generalization (CoDAG), a simple yet effective learning framework that combines domain adaptation and generalization in a complementary manner to achieve three major goals of unsupervised continual domain shift learning: adapting to a current domain, generalizing to unseen domains, and preventing forgetting of previously seen domains. Our approach is model-agnostic, meaning that it is compatible with any existing domain adaptation and generalization algorithms. We evaluate CoDAG on several benchmark datasets and demonstrate that our model outperforms state-of-the-art models in all datasets and evaluation metrics, highlighting its effectiveness and robustness in handling unsupervised continual domain shift learning.

1. Introduction

Machine learning algorithms have found extensive applications in various fields such as image recognition [19, 23], natural language processing [3, 9], and autonomous driving [5, 10]. Typically, these algorithms learn from a training dataset to build a model that performs a target task on new data. The assumption underlying these algorithms is that the training data and the test data are identically and independently distributed (IID) [18], drawn from the same distribution that is characterized by an environment or domain. However, this IID assumption is often not valid in real-world scenarios, as the environment in which the model is applied is more likely to change over time than remain fixed. This implies that the model will encounter new data from various domains over time, and its performance may decline if the data is from a domain that differs significantly from the one it was trained on.

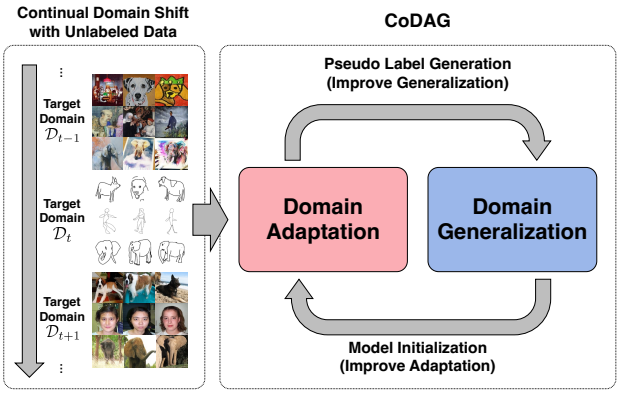

Figure 1. Our proposed Complementary Domain Adaptation and Generalization (CoDAG) framework for unsupervised continual domain shift learning.

To address this issue, two main approaches have been developed: domain generalization and domain adaptation. Domain generalization [60] is a method to enhance a model's ability to generalize to unseen domains by training the model on labeled data from one or more domains, without assuming any prior knowledge of the test environment in which the model will be applied. However, collecting a large volume of labeled data from various domains for a particular task can be challenging in practice. Single-source domain generalization techniques [46] have been developed to address this issue, which rely solely on data from a single domain. Although more practical, these techniques generally have lower generalization abilities compared to multi-source domain generalization techniques.

*Corresponding Author

On the other hand, domain adaptation [61] aims to enhance model performance only on the current target domain and does not prioritize performance on all other domains. In particular, unsupervised domain adaptation (UDA) [44] techniques leverage unlabeled data to adapt the model to a new target domain. However, domain adaptation methods fundamentally suffer from performance degradation on a new target domain before and during the adaptation process due to the lack of inherent mechanism to prepare for unseen domains.

In this paper, we tackle a challenging problem that simulates real-world scenarios where models face continual domain shifts and no labeled data is available for new domains. We refer to this problem as *unsupervised continual domain shift learning*. In this setting, the model must continually adapt to new domains (domain adaptation), while maintaining its generalization ability for upcoming and unseen domains (domain generalization), in an unsupervised manner. However, achieving both objectives simultaneously is not always feasible since they involve related but distinct goals. For instance, if the current target domain is vastly dissimilar from any other domains, none of the optimal solutions for adapting to the current target domain with DA would necessarily result in optimal generalization for performing well on other domains. Similarly, achieving optimal generalization for unseen domains through DG may not result in the best solution for the current target domain. Therefore, to address unsupervised continual domain shift learning, it is necessary to find a solution that resolves this trade-off between domain adaptation and generalization.

To address the trade-off between domain adaptation and generalization, we propose *Complementary Domain Adaptation and Generalization (CoDAG)*, a learning framework that combines domain adaptation and domain generalization in a complementary manner. As shown in Fig. 1, our approach involves training two separate models: one for domain adaptation and the other for domain generalization. We use the domain adaptation model to adapt to the target domain, generating more accurate and reliable pseudo-labels for training the domain generalization model. In turn, the domain generalization model learns more generalized representations across multiple domains and provides the domain adaptation model with initializing parameters, enhancing its adaptability to a new domain. As a result, the domain adaptation and generalization models complement each other in our framework, leading to improved performance for both.

The main contribution of our framework, CoDAG, lies in *the complementary manner in which we leverage existing domain adaptation and domain generalization methods to address unsupervised continual domain shift learning*, a unique and challenging problem that has not been thoroughly explored. We deliberately apply existing methods to our framework, rather than introducing new ones, to underscore that the effectiveness of our framework is due to its complementary structure, not its individual components. Indeed, without requiring any models tailored for the present problem, our framework proved its merit by achieving SoTA performance against all baselines, including the one which is explicitly designed for this setting [37].

Finally, it is important to note that that our work is one of the first attempts to explore the potential synergies between domain adaptation and domain generalization methods. We are breaking new ground by bridging the divide between the disparate fields of domain adaptation and generalization, which were primarily studied independently. This paradigm shift represents not just a novel approach, but one with profound practical implications.

Our contributions can be summarized as follows:

- We introduce a novel framework that combines domain adaptation and generalization models in a complementary manner, resulting in a synergistic process that enhances overall performance.
- Our method consistently outperforms state-of-the-art models across all datasets and metrics, demonstrating superior robustness with the lowest standard deviation across different orders in almost all cases.
- Our approach does not necessitate the use of models designed for the present problem, allowing seamless integration with existing domain adaptation and generalization algorithms for broader applications.

2. Related Work

Domain generalization Domain generalization (DG) is the process of training a model using labeled data from one or multiple domains, with the objective of achieving good generalization performance across unseen domains. Existing DG methods are based on domain-invariant learning [13, 16, 15, 29, 32, 40], meta-learning [2, 11, 26, 28], and data augmentation [52, 67]. To address practical scenarios, single-source DG methods [46, 50, 58, 65, 66] have been proposed, which use labeled data collected from a single domain. However, these techniques require a large amount of labeled data and can suffer from severe catastrophic forgetting when applied to scenarios with continual domain shift.

Unsupervised domain adaptation Unsupervised domain adaptation (UDA) [44] aims to improve the performance of a target model in scenarios where there is a domain shift between the labeled source domain and the unlabeled target domain. UDA methods often achieve distribution alignment through domain invariant feature transformation [34, 39, 43] or feature space alignment [12, 17, 55].

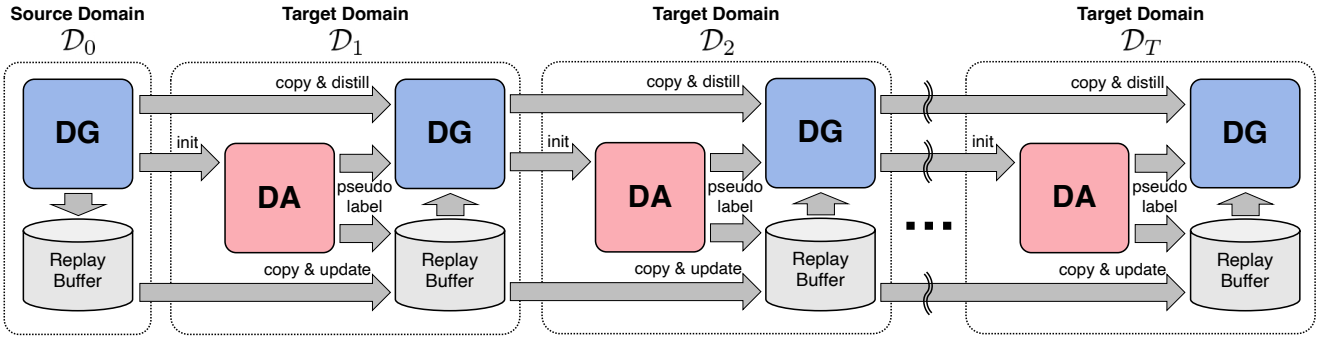

Figure 2. An implementation of CoDAG for unsupervised continual domain shift learning.

Typical domain adaptation techniques generally assume having access to the source data throughout the adaptation process, which is impractical in real-world scenarios. However, source-free domain adaptation (SFDA) methods [35, 36] adapt a model to the unlabeled target domain even when the source dataset is not available during the target adaptation process. Recent works on SFDA [1, 7, 31, 38, 63] have arisen, which use generative models to model the distribution of target data by creating pseudo-label refinement [1, 7], target-style images [31, 38], or variational inference for generating latent source features [63]. Nevertheless, these techniques are not intended to accumulate the knowledge acquired from continually shifting domains.

Continual domain adaptation Continual learning (CL) [8] focuses on avoiding catastrophic forgetting when learning new tasks by using regularization-based methods [53, 64] and replay-based methods [47]. Recent works [4, 51, 56] have adopted the ideas from continual learning to tackle the continual domain adaptation (CDA) problem. These methods involve distilling probability distributions at multiple levels from the previous models to solve the catastrophic forgetting problem [51], using sample replay buffers along with domain adversarial training [4], and utilizing a domain-specific memory buffer for each domain [56]. Despite the use of CDA methods, the model's performance on the target domain is often poor before and during the adaptation process, although it may improve after sufficient adaptation on the new target domain. This can be particularly problematic when there is a significant domain shift. Our complementary framework overcomes this issue by leveraging the DG model's generalization ability to initialize the DA model.

3. Methodology

3.1. Unsupervised continual domain shift learning

We adopt the problem setting introduced by [37] to address the challenge of *unsupervised continual domain shift learning*. Specifically, our approach works with $T+1$ distinct domains $\mathcal{D}_t$ over $t = 0, 1, \ldots, T$ for a given target task, which are sequentially encountered. Here, we tackle the K-way image classification problem as the target task, and the image space $\mathcal{X}$ and the label space $\mathcal{Y}$ are shared across all domains. For the sake of notational simplicity, here we use the notation $\mathcal{D}_t$ to denote both the t-th domain and the dataset sampled from it interchangeably.

In this setting, the first domain $\mathcal{D}_0$ is regarded as the source domain $\mathcal{S}$ that is used to initially learn the target task. Different from other domains, it consists of labeled samples $\mathcal{S} = \mathcal{D}_0 = \{(x_0^{(i)}, y_0^{(i)})\}_{i=1}^{N_0}$, where $x_0^{(i)}$ and $y_0^{(i)}$ denote input data and its corresponding label of the i-th sample and N_0 indicates the total number of samples in the source domain. After a model is initially trained with the source domain, it sequentially encounters a series of target domains $\mathcal{T} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T\}$. In contrast to the source domain, *the target domains do not provide any label information to the model*, as we assume that the model encounters them after the deployment. Moreover, we consider more realistic settings that *samples from previously experienced domains are not fully accessible* but are partially available via a replay buffer with a limited capacity. Therefore, the model has to properly improve its generalization ability in an unsupervised manner by mainly using unlabeled samples $\mathcal{D}_t = \{x_t^{(i)}\}_{i=1}^{N_t}$ from the target domain $\mathcal{D}_t$ at each stage t.

To deal with this problem setting, a model $f(x;\theta)$ is used to map an input $x \in \mathcal{X}$ to its corresponding label $y \in \mathcal{Y}$, where θ represents the set of model parameters to be learned. The model is expected to be initially trained with samples from the source domain $\mathcal{S}$ and then adapted to the target domains $\mathcal{T}$ in a continual manner (*i.e.*, from $\mathcal{D}_1$ to $\mathcal{D}_T$). We denote the model parameters after training on the t-th domain as θ_t^* and evaluate the corresponding model $f(x;\theta_t^*)$ with x from different domains. There exist three main goals for this problem setting. (i) Firstly, we seek to *achieve domain adaptation* for the target domain $\mathcal{D}_t \in \mathcal{T}$. (ii) Secondly, we aim to *achieve domain generalization* for

unseen target domains $\{\mathcal{D}_{t'}\}_{t<t'\leq T}$. (iii) Finally, we aim to *prevent catastrophic forgetting* of knowledge gained from previously seen domains $\{\mathcal{D}_{t'}\}_{0\leq t'<t}$.

Most of the existing models for domain adaptation or generalization, which are able to handle unlabeled data, can be applied for unsupervised continual domain shift learning. A naive approach is to start with the model $f(x;\theta^*_{t-1})$ trained on domain $\mathcal{D}_{t-1}$ and update it with samples from $\mathcal{D}_t$, resulting in a new model $f(x;\theta^*_t)$ that is either adapted to $\mathcal{D}_t$ or generalized for unseen domains $\mathcal{D}_{t'}$ for $t'>t$, depending on its designed purpose. All the baselines used in our experiments reflect this naive approach.

However, achieving both DA and DG objectives in unsupervised continual domain shifts is challenging with a naive single-model approach, given the potential contradiction between adapting to a target domain (DA) and learning domain-invariant features (DG). For instance, when the current target domain is significantly different from any other domains, optimal solutions for domain adaptation may not lead to optimal generalization for other domains. Similarly, using domain generalization to attain optimal generalization for unseen domains does not necessarily yield the best solution for the current target domain.

3.2. CoDAG: Complementary Domain Adaptation and Generalization

To resolve this trade-off between DA and DG, we propose a novel framework called *Complementary Domain Adaptation and Generalization (CoDAG)*. CoDAG is highly effective yet remarkably simple. Unlike other approaches that require complicated architectures or learning methods, our CoDAG framework relies on a straightforward and intuitive idea: maintaining two separate models for DA and DG. By having two distinct models optimized for their respective goals, the CoDAG framework eliminates the need for complicated techniques that attempt to balance DA and DG within a single model. Instead, our framework facilitates a synergistic relationship between the two models, where they complement each other, leading to improved performance for both.

To carry out this framework, DA and DG are conducted in an interleaved order with two separate models for DA and DG, and we make each of them to be dependent on the other, as shown in Fig. 2. Specifically, we duplicate the model f to use one for DA, denoted as f_{DA}, and the other for DG, denoted as f_{DG}, with each model trained with DA and DG algorithms, respectively. This setting allows the models to effectively exchange knowledge and continually improve the performance for both.

As domain shifts occur continuously and there is no labeled information available from target domains, nor access to data from previously encountered domains, we have implemented a source-free unsupervised domain adaptation algorithm for our DA model. There are several options for this, but we choose a simple pseudo-labeling [25] along with regularization from SHOT [35] due to its simplicity. For the DG model, we simply apply an empirical risk minimization (ERM) method along with a RandMix augmentation technique [37] that generates diversified data, which boost the DG model's ability to generalize to unseen domains. However, note that our proposed framework is model-agnostic, in the sense that it can be applied to any DA or DG algorithms that are suitable for the given problem setting. The implementation details of auxiliary methods and algorithms employed in this paper are provided in the Supplementary Material.

Source domain training To initially learn a given target task, we train the DG model from scratch with the set of labeled samples from the source domain $\mathcal{S}$ $(=\mathcal{D}_0)$. This stage can be approached by a *single-source* domain generalization method. In that way, we simply minimize the following cross-entropy loss over the source domain data with enhanced data augmentation,

$$\mathcal{L}_{\text{DG},0}(\mathcal{D}_0)=\mathbb{E}_{(x,y)\in\mathcal{D}_0}\left[\mathcal{L}^{\text{ce}}(f_{\text{DG}}(R(x);\theta_{\text{DG},0}),y)\right], \quad (1)$$

where $\theta_{\text{DG},0}$ is the DG model parameters, $\mathcal{L}^{\text{ce}}$ is the cross-entropy loss for a classification setting, and R represents the RandMix augmentation [37].

Generalized initialization with DG for DA In general, unsupervised source-free domain adaptation approaches involve initially training a source model with the source domain data and further updating the source model with unlabeled data from a new target domain. However, in our proposed framework, the DA model utilizes the parameters of the previous DG model for its initialization. This allows the DA model to leverage the DG model's generalization ability to learn domain-invariant features and reduce domain-specific factors. As a result, we achieve efficient adaptation to a new target domain, even when there is a large gap between previously experienced domains and the new target domain (see Section 4.2 for experimental results).

To apply this approach for the current target domain $\mathcal{D}_t$, we first initialize the DA model f_{DA} with the parameters of the DG model trained with previously experienced domains, or $\theta^*_{\text{DG},t-1}$, treating it as a source model. Then, we freeze the classifier head in f_{DA} and only update the feature extractor part of it using information maximization and self-supervised pseudo-labeling with data from the current target domain $\mathcal{D}_t$. Accordingly, the loss to adapt to $\mathcal{D}_t$ is written as,

$$\mathcal{L}_{\text{DA},t}(\mathcal{D}_t)=\mathbb{E}_{x\in\mathcal{D}_t}\left[\mathcal{L}^{\text{shot}}(f_{\text{DA}}(x;\theta_{\text{DA},t}))\right], \quad (2)$$

where $\theta_{\text{DA},t}$ is the parameters of the DA model initialized with the optimal parameters of the DG model trained on the

previous domain $\mathcal{D}_{t-1}$, or $\theta^*_{\text{DG},t-1}$, and $\mathcal{L}^{\text{shot}}$ is the cross-entropy loss with pseudo-labels on target predictions along with regularization from SHOT [35].

Pseudo-label generation with DA for DG In our proposed framework, we simply use an empirical risk minimization (ERM) method along with an enhanced data augmentation method for training the DG model. Although labeled samples are necessary for this process, none of the target domains $\mathcal{T}$ provide any labels. Thus, to make use of unlabeled samples from the current target domain $\mathcal{D}_t$, we adopt a pseudo-label generation strategy [25] based on the highest prediction confidence of the DA model adapted to $\mathcal{D}_t$. This involves applying the DA model to the unlabeled samples from $\mathcal{D}_t$ to generate pseudo-labels, which are then used as training labels for the DG model on $\mathcal{D}_t$.

Specifically, for each unlabeled sample x, we compute its pseudo-label $\hat{y}_t(x)$ as follows:

$$\hat{y}_t(x) = \underset{k}{\operatorname{argmax}}\, \delta_k(f_{\text{DA}}(x; \theta^*_{\text{DA},t})), \tag{3}$$

where $\theta^*_{\text{DA},t}$ is obtained by optimizing $\mathcal{L}_{\text{DA},t}$ in Eq. 2 and $\delta_k(\cdot)$ is the k-th element of a softmax output. By using the resulting pseudo-labels as training labels for the DG model on $\mathcal{D}_t$, we construct a pseudo-labeled dataset $\hat{\mathcal{D}}_t = \{(x_t^{(i)}, \hat{y}_t(x_t^{(i)}))\}_{i=1}^{N_t}$. We then update the DG model with ERM in the same way as the source training in Eq. 1 with the following loss:

$$\mathcal{L}^{\text{erm}}_{\text{DG},t}(\hat{\mathcal{D}}_t) = \mathbb{E}_{(x,y)\in\hat{\mathcal{D}}_t}\left[\mathcal{L}^{\text{ce}}(f_{\text{DG}}(R(x); \theta_{\text{DG},t}), y)\right], \tag{4}$$

where $\theta_{\text{DG},t}$ is initialized with the optimal parameters of the DG model trained on the previous domain $\mathcal{D}_{t-1}$, or $\theta^*_{\text{DG},t-1}$.

Learning from noisy labels As the DA model is optimized to adapt to the current domain $\mathcal{D}_t$, we assume that it can generate high-quality pseudo-labels. However, some of the labels may still contain errors due to the imperfectness of the DA method. Unfortunately, the errors in pseudo-labels (*i.e.*, noisy labels [54]) can negatively affect the performance of the DG model. To alleviate this problem, we use an algorithm that can properly handle noisy labels to prevent the performance degradation of the DG model.

In this paper, we adopt an algorithm called Selective Negative Learning and Positive Learning (SelNLPL) [22] to reduce the risk of overfitting to noisy labels and improve performance of the DG model. We show the effectiveness of this approach in Sec. 4.2. Note that our approach offers greater flexibility that is not restricted to SelNLPL but rather can leverage any label noise-resilient methods [33, 48, 57].

Forgetting alleviation Forgetting alleviation is crucial in unsupervised continual domain shift learning, where limited access to data from previous domains makes it challenging to maintain performance on previous tasks. When encountering a new target domain, the performance of the DG model on previous domains tends to degrade, which is commonly referred to as catastrophic forgetting in continual learning.

To address this issue, we add a simple distillation loss term [20] $\mathcal{L}^{\text{distill}}_{\text{DG},t}$ to the loss of the DG model in Eq. 4 to ensure that the model retains the knowledge gained from previous domains $\{\mathcal{D}_{t'}\}_{0\le t'<t}$ while learning from the current target domain $\mathcal{D}_t$, given by

$$\mathcal{L}^{\text{distill}}_{\text{DG},t}(\mathcal{D}_t) = \mathbb{E}_{x\in\mathcal{D}_t}\left[\mathcal{L}^{\text{kl}}(q_t(x)||p_t(x))\right], \tag{5}$$

where $\mathcal{L}^{\text{kl}}$ represents the KL divergence loss, and $q_t(x) = \delta(f_{\text{DG}}(R(x); \theta^*_{\text{DG},t-1}))$ and $p_t(x) = \delta(f_{\text{DG}}(R(x); \theta_{\text{DG},t}))$ are the predicted softmax probabilites from the previous and current DG models, respectively.

Another method we employ to prevent catastrophic forgetting is a replay buffer $\mathcal{M}_t$ of size $M \ll N_t$ [49], which contains selected samples from previously experienced domains $\{\mathcal{D}_{t'}\}_{0\le t'<t}$. We build the replay buffer based on the iCaRL approach [37, 47]. By using a replay buffer, along with the samples from $\mathcal{D}_t$, additional M selected samples from $\mathcal{D}_0 \cup \{\hat{\mathcal{D}}_{t'}\}_{1\le t'<t}$ are available for training the DG model on $\mathcal{D}_t$. This allows the DG model to not only improve its generalization ability but also prevent catastrophic forgetting. Then, our final loss to update the DG model on the current target domain $\mathcal{D}_t$ is given by,

$$\mathcal{L}_{\text{DG},t}(\tilde{\mathcal{D}}_t) = \mathcal{L}^{\text{erm}}_{\text{DG},t}(\tilde{\mathcal{D}}_t) + \alpha \cdot \mathcal{L}^{\text{distill}}_{\text{DG},t}(\tilde{\mathcal{D}}_t), \tag{6}$$

where $\tilde{\mathcal{D}}_t = \hat{\mathcal{D}}_t \cup \mathcal{M}_t$ represents $N_t + M$ samples available from the t-th domain with a replay buffer and α is a balancing hyperparameter.

Our findings suggest that utilizing a replay buffer leads to a significant improvement in performance, especially when dealing with previous domains. However, even without a replay buffer, our experiment shows that our model remains competitive against state-of-the-art models that are equipped with replay buffers (see Sec. 4.2 for experimental results).

4. Experiments

To validate the effectiveness of our framework, we compare our proposed framework, CoDAG, against state-of-the-art methods on three benchmark datasets: (i) PACS [27], (ii) Digits-five [14, 21, 24, 41], and (iii) DomainNet [45]. For a fair comparison, we adhere to the experimental setup from [37] in the following sections.

Datasets PACS consists of 4 distinct domains with 7 classes, including Photo (P), Art painting (A), Cartoon

Table 1. Comparison of the performance on the PACS, Digits-five, and DomainNet datasets for different state-of-art methods in TDA, TDG, FA, and All. The results are averaged over 10 different orders from each dataset. The results of the baseline models are referenced from [37]. The best results are highlighted in bold. Our method outperforms all the baselines across all datasets and evaluation metrics tested.

Dataset	Metric	Comparison Baselines (*w/ Replay Buffer*)								Ours
		SHOT+ [36, 37]	SHOT++ [36]	Tent [59]	AdaCon [6]	EATA [42]	L2D [62]	PDEN [30]	RaTP [37]	CoDAG
PACS	TDA	81.9 ±9.2	84.4 ±8.0	78.7 ±6.9	79.9 ±5.9	80.3 ±7.1	78.8 ±5.6	77.8 ±5.2	84.7 ±5.1	**87.6** ±4.0
	TDG	54.9 ±13.1	56.0 ±10.9	65.8 ±11.5	65.2 ±10.5	64.1 ±12.1	65.8 ±9.6	64.4 ±9.8	70.6 ±9.1	**72.2** ±8.3
	FA	74.9 ±8.1	83.0 ±4.0	81.0 ±6.2	81.6 ±5.9	82.6 ±7.0	77.6 ±4.6	76.3 ±4.0	83.9 ±4.7	**88.8** ±3.0
	All	70.6 ±9.2	74.5 ±5.7	75.2 ±7.8	75.6 ±7.1	75.7 ±8.6	74.1 ±6.2	72.9 ±5.9	79.7 ±5.7	**82.9** ±4.8
Digits-five	TDA	78.6 ±13.2	81.3 ±14.0	68.7 ±11.0	71.6 ±9.2	72.0 ±9.8	84.3 ±5.4	82.3 ±5.8	88.7 ±1.8	**92.7** ±1.7
	TDG	61.0 ±14.9	62.3 ±13.8	64.0 ±13.6	63.3 ±13.1	64.0 ±12.9	70.9 ±6.8	69.7 ±7.0	76.8 ±3.9	**77.4** ±4.3
	FA	58.2 ±14.9	64.5 ±13.3	66.1 ±15.7	72.2 ±11.2	73.0 ±10.9	76.5 ±3.8	74.0 ±4.0	85.0 ±2.2	**87.1** ±2.1
	All	65.9 ±13.5	69.4 ±12.9	66.2 ±13.3	69.1 ±11.0	69.6 ±10.9	77.2 ±4.8	75.3 ±5.1	83.5 ±2.1	**85.7** ±2.2
DomainNet	TDA	66.0 ±8.8	66.9 ±8.7	53.6 ±13.2	62.2 ±7.7	62.5 ±7.3	56.2 ±6.2	55.6 ±6.6	65.4 ±5.1	**71.0** ±5.7
	TDG	47.3 ±11.0	48.1 ±10.7	47.7 ±11.0	51.3 ±10.0	52.1 ±9.9	50.7 ±9.1	49.3 ±9.1	55.2 ±7.4	**56.2** ±7.2
	FA	58.5 ±8.3	66.9 ±6.0	56.1 ±14.5	61.8 ±9.0	62.8 ±8.8	52.2 ±9.4	50.2 ±9.5	63.5 ±6.6	**70.9** ±6.6
	All	57.3 ±8.9	60.6 ±8.0	52.5 ±12.4	58.4 ±8.6	59.1 ±8.3	53.0 ±7.6	51.7 ±7.8	61.4 ±6.0	**66.0** ±6.2

(C), and Sketch (S). Digits-five contains 5 different domains with 10 classes, 0 to 9, including MNIST (MT) [24], MNIST-M (MM) [14], SVHN (SN) [41], SYN-D (SD) [14] and USPS (US) [21]. DomainNet is the most challenging dataset, which includes Quickdraw (Qu), Clipart (Cl), Painting (Pa), Infograph (In), Sketch (Sk) and Real (Re). DomainNet has an imbalance in class distribution, where some domains have limited images for certain classes. To address this issue, a subset of DomainNet is used by selecting the top 10 classes with most images in the whole dataset.

Experimental settings In the source domain, 80% of the data is randomly assigned as a training set, and the remaining 20% as a testing set. In target domains, all data is used for training and testing as we assume any label information is not available. The experiments are repeated three times using different seeds (2022, 2023, 2024), and the average performance is reported. For Digits-five, we use DTN [35] as a feature extractor, while for both PACS and DomainNet, we employ ResNet-50 [19]. The SGD optimizer is used with a batch size of 64 for all experiments. The size of replay buffer is set to 200 for all datasets. See the Supplementary Material for more details on the experimental settings and network architectures.

Evaluation metrics (i) TDA: *Target domain adaptation* for each domain is the performance measured on the domain right after its training stage. The TDA of the t-th domain for $t = 0, \ldots, T$ is

$$\text{TDA}_t = \mathcal{A}(f(x; \theta_t^*), \mathcal{D}_t), \tag{7}$$

where $\mathcal{A}(f(x; \theta_{t'}^*), \mathcal{D}_t)$ represents the test accuracy on the domain $\mathcal{D}_t$ with the model $f(x; \theta_{t'}^*)$ obtained after training on the domain $\mathcal{D}_{t'}$. (ii) TDG: *Target domain generalization* for each domain is evaluated by the average performance on the domain prior to its training stage. The TDG of the t-th

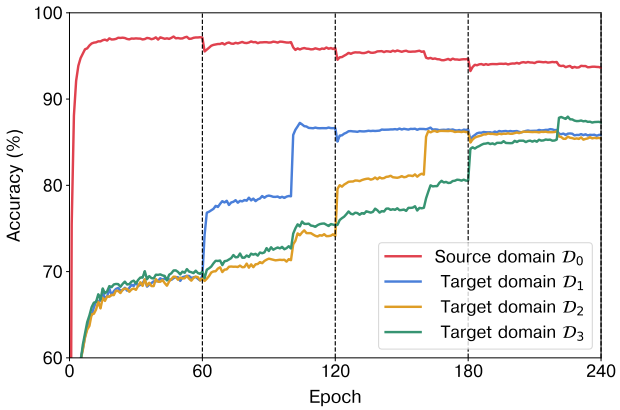

Figure 3. The training curves of the DG model which depict the model's performance on their respective domains at every training epoch. The domain shift occurs every 60 epochs, starting from the source domain and continuing until the third target domain. The results are averaged over 10 different orders from the PACS dataset.

domain for $t = 1, \ldots, T$ is given by

$$\text{TDG}_t = \frac{1}{t} \sum_{t'=0}^{t-1} \mathcal{A}(f(x; \theta^*_{t'}), \mathcal{D}_t). \tag{8}$$

(iii) FA: *Forgetting alleviation* for each domain is evaluated by the average performance on the domain after the model has been trained on subsequent domains. The FA of the t-th domain for $t = 0, \ldots, T-1$ can be written as

$$\text{FA}_t = \frac{1}{T-t} \sum_{t'=t+1}^{T} \mathcal{A}(f(x; \theta^*_{t'}), \mathcal{D}_t). \tag{9}$$

For comparative analysis between different models, we use the averaged value of each metric over all domains for which the metric is defined. Additionally, we define the average of all the metrics as a composite score (All) to evaluate the overall performance of the models. To properly evaluate our proposed framework, CoDAG, we use the DA model f_{DA} to evaluate TDA, while the DG model f_{DG} is used to evaluate TDG and FA.

4.1. Effectiveness of the CoDAG framework

Comparing with state-of-the-art To compare the performance of our model with state-of-the-art methods, we use the experiment results of different methods reported in [37] as baselines. These methods include several state-of-the-art models from Source-Free DA (SHOT+ [36, 37] and SHOT++ [36]), Test-Time/Online DA (Tent [59], AdaCon [6], and EATA [42]), Single DG (L2D [62] and PDEN [30]), as well as Continual DG (RaTP [37]). To ensure consistency and fairness, we adopt the same backbone feature extractor with the baseline methods. In the present experiments, all baseline methods are equipped with the replay buffer of size 200.

Table 2. Evaluation of two distinct initialization methods for the DA model with performance averaged across 10 different orders from the PACS dataset.

Method	TDA	TDG	FA	All
Initialization w/ the DG model	87.6 ±4.0	72.2 ±8.3	88.8 ±3.0	82.9 ±4.8
Initialization w/ the DA model	83.8 ±4.0	71.7 ±8.4	86.6 ±4.0	80.7 ±4.8
Diff.	+3.8	+0.5	+2.2	+2.2

We present an evaluation of our CoDAG framework against the baseline models on the three datasets (PACS, Digits-five, and DomainNet) using the four evaluation metrics (TDA, TDG, FA, and All). The results are averaged over 10 different orders from each dataset to ensure the robustness of our findings. The overall results presented in Table 1 demonstrate that our method consistently outperforms all the baseline models across all datasets and evaluation metrics.

Compared to RaTP [37], a continual domain generalization method that achieves the best performance among the baselines in most cases, our method demonstrates significantly higher performance in TDA and FA. This highlights the effectiveness of our domain adaptation stage based on the generalized initialization using the DG model. The more accurate pseudo-labels generated by the DA model also contribute to the DG model's superior performance in FA.

On the DomainNet dataset, which is the most challenging among our benchmark datasets, the DA-specialized model SHOT++ [36] outperforms RaTP in terms of TDA and FA. However, our CoDAG consistently outperforms any DA-specialized models in terms of TDA and FA, while maintaining the improved generalization performance in terms of TDG.

Furthermore, Table 1 shows that our method acheives the lowest standard deviation across ten different orders in nearly all cases. Notably, even our lower bound performance $(\mu - \sigma)$ surpasses the upper bound performance $(\mu + \sigma)$ of other baselines in many cases. These findings demonstrate the robustness of our CoDAG framework in addressing the challenges of unsupervised continual domain shift learning.

Training curves In Fig. 3, we display the training curves of the DG model, each of which represents the accuracy of the model on its corresponding domain at different training

Table 3. The ablation study of SelNLPL conducted for all possible pairs of two domains ($\mathcal{D}_0 \rightarrow \mathcal{D}_1$) from the PACS dataset. TDG was measured by the average performance on two unseen domains ($\mathcal{D}_2$ and $\mathcal{D}_3$), while FA was measured by the performance on the source domain ($\mathcal{D}_0$). Diff. denotes the result obtained by subtracting the performance without (w/o) SelNLPL from the performance with (w/) SelNLPL.

Metric	Method	P→A	P→C	P→S	A→P	A→C	A→S	C→P	C→A	C→S	S→P	S→A	S→C	Avg.
TDG	w/ SelNLPL	58.3	74.6	58.4	59.8	85.3	77.7	79.8	86.8	78.1	69.9	84.3	85.3	74.9
	w/o SelNLPL	57.4	74.2	54.3	58.0	84.9	71.2	78.6	86.6	74.2	68.3	83.4	85.0	73.0
	Diff.	+0.9	+0.4	+4.1	+1.8	+0.4	+6.5	+1.2	+0.2	+3.9	+1.6	+0.9	+0.3	+1.8
FA	w/ SelNLPL	98.5	98.2	95.7	94.1	93.2	83.0	90.4	92.6	85.8	85.7	91.2	89.7	91.5
	w/o SelNLPL	98.5	97.2	93.0	93.2	92.6	76.7	89.4	91.3	84.7	85.1	89.3	88.4	89.9
	Diff.	0.0	+1.0	+2.7	+0.9	+0.6	+6.3	+1.0	+1.3	+1.1	+0.6	+1.9	+1.3	+1.6

stages. We observe that the model's performance on unseen domains gradually increases as training progresses, which illustrates the model's continually improving generalization ability. Moreover, Fig. 3 suggests that the model is capable of avoiding catastrophic forgetting, as its performance on previously encountered domains remains relatively stable even after domain shifts.

4.2. Further analysis

In this section, we perform additional analyses to investigate the key components of our method using the PACS dataset under different experimental settings.

Effectiveness of generalized initialization for DA To evaluate the effectiveness of the generalized initialization approach for DA, we compare model performance between two different approaches for initializing $\theta_{\text{DA},t}$ of the DA model on $\mathcal{D}_t$: (1) initializing with the previous DG model using $\theta^*_{\text{DG},t-1}$ and (2) initializing with the previous DA model using $\theta^*_{\text{DA},t-1}$. The results presented in Table 2 show that initializing with the DG model significantly improves the performance in TDA. This exhibits the effectiveness of the generalized initialization, which enables the DA model to adapt more efficiently to a new domain, compared to relying on parameters specifically adapted to the previous domain. Furthermore, we observe that the improved TDA of the DA model has a positive impact on the performance of the DG model in FA and TDG by providing more accurate pseudo-labels.

Ablation study of SelNLPL To understand SelNLPL's contribution, we conduct experiments with and without SelNLPL in the training of our model for pairs of two different domains ($\mathcal{D}_0 \rightarrow \mathcal{D}_1$). We then compare the resulting changes in TDG and FA. To isolate the effect of SelNLPL, we remove the replay buffer. The experiment results in Table 3 demonstrate that SelNLPL can improve model performance in both TDG and FA, which underscores the effectiveness of noise-resilient methods to amplify the complementary effect of the DA model on the DG model within our framework.

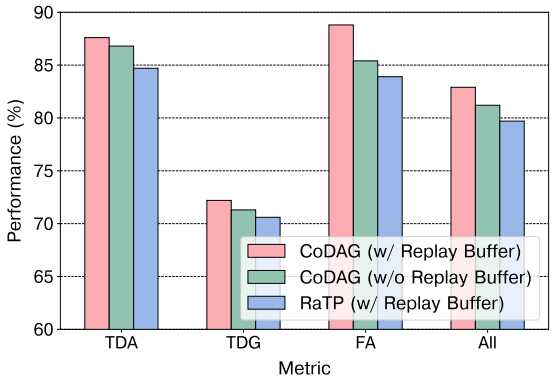

Figure 4. The ablation study of replay buffer with averaged performance across 10 different orders from the PACS dataset.

Ablation study of replay buffer We perform ablation studies to assess the effectiveness of utilizing a replay buffer in CoDAG, by comparing the performance of the model with and without the buffer. The results presented in Fig. 4 clearly display that the absence of the buffer leads to a degradation in performance across all metrics, with the metric for FA being the most adversely affected. This indicates that the replay buffer plays a crucial role in preventing catastrophic forgetting.

Furthermore, our findings suggest that the replay buffer enhances the model's ability to generalize by continually exposing it to multiple domains, thereby boosting its performance in TDG. This, in turn, leads to an improvement in TDA as the model can adapt to a new domain with more generalized initialization.

However, despite the decrease in performance after the removal of the buffer, our method *without replay buffer* still outperforms all other baselines *with replay buffer*, further confirming the effectiveness of our framework.

5. Conclusion

In this paper, we propose a learning framework that combines domain adaptation and generalization models in a complementary manner, which effectively addresses the challenge of unsupervised continual domain shift learning. Our method outperforms state-of-the-art performance across different datasets and evaluation metrics. It also achieves competitive results even without a replay buffer, demonstrating its effectiveness and robustness in real-world scenarios.

Our approach is model-agnostic, meaning it can be used with any domain adaptation and generalization algorithms suitable for a given problem. We envision extending our method to complex scenarios beyond a pre-defined set of domains, dynamically discovering domain shifts and adapting to new domains. This can increase its applicability to a broader range of scenarios, with potential contributions to practical applications in computer vision and other fields.

Acknowledgement

This work was supported by the Hyundai Motor Chung Mong-Koo Foundation, the New Faculty Startup Fund from Seoul National University, and IITP (RS-2023-00232046).

References

[1] Waqar Ahmed, Pietro Morerio, and Vittorio Murino. Adaptive pseudo-label refinement by negative ensemble learning for source-free unsupervised domain adaptation. *arXiv preprint arXiv:2103.15973*, 2021.

[2] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018.

[3] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

[4] Andreea Bobu, Eric Tzeng, Judy Hoffman, and Trevor Darrell. Adapting to continuously shifting domains. 2018.

[5] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[6] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022.

[7] Weijie Chen, Luojun Lin, Shicai Yang, Di Xie, Shiliang Pu, and Yueting Zhuang. Self-supervised noisy label learning for source-free unsupervised domain adaptation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10185–10192. IEEE, 2022.

[8] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[11] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019.

[12] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013.

[13] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 87–97, 2016.

[14] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[15] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016.

[16] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015.

[17] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Unsupervised adaptation across domain shifts by generating intermediate data representations. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2288–2302, 2013.

[18] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society.

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[21] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

[22] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 101–110, 2019.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[25] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[26] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[27] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

[28] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019.

[29] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

[30] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2021.

[31] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9641–9650, 2020.

[32] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 624–639, 2018.

[33] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE international conference on computer vision*, pages 1910–1918, 2017.

[34] Jian Liang, Ran He, Zhenan Sun, and Tieniu Tan. Aggregating randomized clustering-promoting invariant projections for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 41(5):1027–1042, 2018.

[35] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[36] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8602–8617, 2021.

[37] Chenxi Liu, Lixu Wang, Lingjuan Lyu, Chen Sun, Xiao Wang, and Qi Zhu. Deja vu: Continual model generalization for unseen domains. In *The Eleventh International Conference on Learning Representations*, 2023.

[38] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021.

[39] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2200–2207, 2013.

[40] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017.

[41] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[42] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022.

[43] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210, 2010.

[44] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[45] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[46] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020.

[47] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[48] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An

uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021.

[49] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[50] Eduardo Romera, Luis M Bergasa, Jose M Alvarez, and Mohan Trivedi. Train here, deploy there: Robust segmentation in unseen domains. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1828–1833. IEEE, 2018.

[51] Antoine Saporta, Arthur Douillard, Tuan-Hung Vu, Patrick Pérez, and Matthieu Cord. Multi-head distillation for continual unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3751–3760, 2022.

[52] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745*, 2018.

[53] Daniel L Silver and Robert E Mercer. The task rehearsal method of life-long learning: Overcoming impoverished data. In *Advances in Artificial Intelligence: 15th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2002 Calgary, Canada, May 27–29, 2002 Proceedings 15*, pages 90–101. Springer, 2002.

[54] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[55] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[56] Shixiang Tang, Peng Su, Dapeng Chen, and Wanli Ouyang. Gradient regularized contrastive learning for continual domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2665–2673, 2021.

[57] Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. *Advances in neural information processing systems*, 28, 2015.

[58] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

[59] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[60] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[61] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[62] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 834–843, 2021.

[63] Hao-Wei Yeh, Baoyao Yang, Pong C Yuen, and Tatsuya Harada. Sofa: Source-data-free feature alignment for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 474–483, 2021.

[64] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.

[65] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, 2020.

[66] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *Advances in Neural Information Processing Systems*, 33:14435–14447, 2020.

[67] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13025–13032, 2020.