GAL: Global-Anchored Learning for Memory-Free Continual Domain Shift Learning

1st Shenghao Lyu
Graduate School of Informatics
Kyoto University
Kyoto, Japan
lyu@ml.ist.i.kyoto-u.ac.jp

2nd Xiaofeng Lin
Graduate School of Informatics
Kyoto University
Kyoto, Japan
lxf@ml.ist.i.kyoto-u.ac.jp

3rd Hisashi Kashima
Graduate School of Informatics
Kyoto University
Kyoto, Japan
kashima@i.kyoto-u.ac.jp

***Abstract*—Deep learning models deployed in dynamic environments often face a challenge of acquiring knowledge from data in continually shifting unlabeled target domains, which is known as Unsupervised Continual Domain Shift Learning. Existing methods typically require a replay buffer that stores past domain data to assist learning of the current target domain. However, this makes their models risky and unsuitable for privacy-sensitive and scalable environments. In this paper, we consider a more practical setting where data from target domains cannot be retained or shared. To address this challenge, we propose a simple yet effective framework named Global-Anchored Learning (GAL). GAL consists of a core module named Anchored Pre-Learning (APL), and a downstream module named Clean Sample Mining (CSM). APL introduces a pre-learning mechanism that captures target-specific knowledge while staying consistent with existing global knowledge. The acquired knowledge is then aggregated to benefit global optimum. CSM further boosts local adaptation through confident sample selection guided by the knowledge from pre-learning. Experimental results on multiple benchmark datasets demonstrate that GAL achieves state-of-the-art (SOTA) or highly competitive performance compared to replay-based methods.**

***Index Terms*—Continual Domain Shift Learning, Domain Adaptation, Domain Generalization.**

I. Introduction

Deep learning models have demonstrated remarkable success in many fields under the independent and identically distributed (IID) assumption [1]. However, this assumption rarely holds in real-world scenarios, where dynamic environments cause test data to originate from continually shifting target domains. This discrepancy substantially limits their practical deployment.

To address this challenge, unsupervised Continual Domain Shift Learning (CDSL) [2] have been proposed to enable models to deal with randomly shifting target domains and continually acquire knowledge from them. Unsupervised CDSL simultaneously tackles three key objectives: **target domain generalization (TDG)**, which aims to enhance performance on unseen or unlearned target domains; **target domain adaptation (TDA)**, which focuses on improving performance on the current target domain; and **forgetting alleviation (FA)**, which seeks to maintain performance on previously encountered domains throughout continual learning.

Although several methods [2], [3] are designed to achieve these objectives, they typically overlook critical constraints in real-world deployment. These methods commonly rely on a replay buffer, which stores data from previously seen domains and is jointly used with current target domain data in centralized training to achieve generalization and alleviate forgetting. Depending on its design, a replay buffer introduces different limitations: When its size is fixed as commonly practiced, it restricts the number of domains the model can effectively learn from. Alternatively, allowing the buffer to grow with each new domain leads to escalating data transmission costs. Regardless of the strategy, both approaches ultimately compromise scalability. More fundamentally, reliance on a replay buffer inherently raises privacy concerns, as it involves storing and sharing data among target domains, which conflicts with the privacy-preserving requirements of modern distributed environments.

In this paper, we address a practical yet underexplored unsupervised CDSL problem where data from seen domains cannot be stored or shared. Unlike the conventional unsupervised CDSL setting, the model has access only to the current target domain during training and cannot revisit data from previously seen domains. This setting not only preserves the privacy in each domain but also eliminates the need to retain or transmit previous data, which in turn improves scalability and extends the applicability of unsupervised CDSL to more realistic, privacy-sensitive environments.

To address this challenge, we propose Global-Anchored Learning (GAL), a replay-free framework for unsupervised CDSL. As illustrated in Fig. 1, GAL is built around its core module, Anchored Pre-Learning (APL), which is responsible for training a pre-learning model that acquires knowledge from the current target domain under the guidance of the existing global knowledge. APL consists of two closely cooperating submodules: Generalizable Shift Learning (GSL) and Global Weight Injection (GWI). GSL defines the fundamental objective of the pre-learning model, which is to learn domain-specific knowledge from the current target domain while retaining a certain level of generalization. Meanwhile, GWI continuously aligns the parameters of the pre-learning model with those of the current global model, preventing the pre-learning model from excessively drifting away from the global optimum established by previously seen domains. This design not only avoids catastrophic forgetting when integrating the

knowledge from the pre-learning model into the global model but also reduces the risk of learning spurious representations. As a result, the global model can approach a new global optimum after knowledge aggregation, thereby improving its overall generalization performance. In addtition, the pre-learning model does not solely serve the global model. To further enhance the adaptability of the framework, a downstream module, Clean Sample Mining (CSM), leverages the confidence scores predicted by the pre-learning model to partition the current domain samples into noisy and clean subsets using a two-component Gaussian Mixture Model (GMM). By isolating a cleaner subset with higher label accuracy, CSM enables the model to achieve improved adaptation performance. Extensive experiments on three domain-shift benchmarks demonstrate that GAL achieves state-of-the-art performance in both domain adaptation and domain generalization without relying on previous data, while maintaining competitive forgetting alleviation performance to existing methods. The source code is publicly available at https://github.com/ShenghaoLyu/GAL.

Our main contributions are summarized as follows:

- We tackle the challenges of unsupervised CDSL in privacy-sensitive and scalable scenarios by proposing GAL, which ensures feasibility without storing or sharing previous domain data.
- We introduce APL with GSL and GWI submodules to balance the learning of target-specific knowledge and maintaining existing global knowledge.
- Extensive experiments on three popular benchmark datasets demonstrate that GAL achieves competitive and even state-of-the-art performance compared to replay-based methods, which validates GAL as a reliable and effective approach.

II. Related Work

We first introduce Domain Adaptation (DA) and Domain Generalization (DG), two pioneering fields that address domain shifts caused by non-IID data, and discuss their limitations in dynamic and randomly-shifted environments. We then present unsupervised Continual Domain Shift Learning (CDSL), along with existing methods and their inherent constraints.

a) Domain Adaptation: Domain Adaptation (DA) [4] aims to adapt a model, initially trained on one or more source domains, to a target domain with a different data distribution. In the traditional DA setting, labeled source data and partially labeled target data are both accessible during training. Later studies have sought to relax these constraints. Unsupervised Domain Adaptation (UDA) [5], [6] focuses on adapting the model without any labeled data from the target domain, and has become a widely adopted setting. Source-Free Domain Adaptation (SFDA) [7], [8] further extends the applicability of DA by considering scenarios where source data is no longer accessible during adaptation. Continual Domain Adaptation (CDA) [9], [10] focuses on sequentially adapting the model to a series of incoming target domains while maintaining its performance on previously seen domains. However, DA is applicable only when the target domain data is accessible and fully leveraged during training, making it unsuitable for handling continual domain shifts in dynamic environments.

b) Domain Generalization: Domain Generalization (DG) [11] aims to enable models to generalize to out-of-distribution (OOD) data by training on source domains. The key idea is to learn domain-invariant representations that can generalize to unseen target domains. Existing methods are mainly based on data augmentation [12], [13], domain alignment [14]–[16], and meta-learning [17], [18], and typically require access to multiple source domains. Although recent studies on single-source DG [19], [20] and source-free DG [21], [22] have attempted to relax this requirement, DG methods still rely heavily on large amounts of labeled data and are trained in an offline manner. Once the training is completed, their generalization capability becomes fixed and cannot be further enhanced. This limits their ability to handle only a finite degree of domain shift, making them unsuitable for learning and optimizing under continual domain shifts.

c) Unsupervised Continual Domain Shift Learning: Unsupervised Continual Domain Shift Learning (CDSL) [2] focuses on a more practical setting, where the environment is dynamic and the distribution of target domains undergoes continual shifts. The model is not only required to adapt to each incoming unlabeled target domain with unpredictable shifts, but also to learn transferable knowledge that enhances its generalization to future unseen domains. Unsupervised CDSL simultaneously targets three objectives: (1) target domain generalization, which aims to continuously enhance the generalization ability of the model to unseen target domains; (2) target domain adaptation, which ensures strong performance on the current target domain after adaptation; (3) forgetting alleviation, which maintains performance on previously learned domains during the continual learning process. [2] maintain a set of prototypes aligned across domains via contrastive learning for domain generalization, while employing a pseudo-labeling strategy for domain adaptation. [3] decouple the DA and DG models and design them to function in a complementary manner. Despite adopting different learning strategies, these methods require a memory buffer to store data from previous domains, which is used to retrain to alleviate catastrophic forgetting and enhance domain generalization. In addition, a recent study [23] further utilizes the stored samples to directly assist inference. Such reliance on stored data limits their applicability in privacy-sensitive or scalable environments. In this paper, we proposed GAL to tackle the problem of unsupervised CDSL without the assistance of a memory buffer.

d) Continual Learning: Continual learning (CL) [24] aims to mitigate catastrophic forgetting over a sequence of tasks. While some recent CL methods [25], [26] avoid explicit memory buffers, they mainly focus on class-incremental learning within a single domain, where the label space expands over time. In contrast, CDSL considers a fixed task and label space under evolving input distributions, leading to fundamentally different modeling objectives.

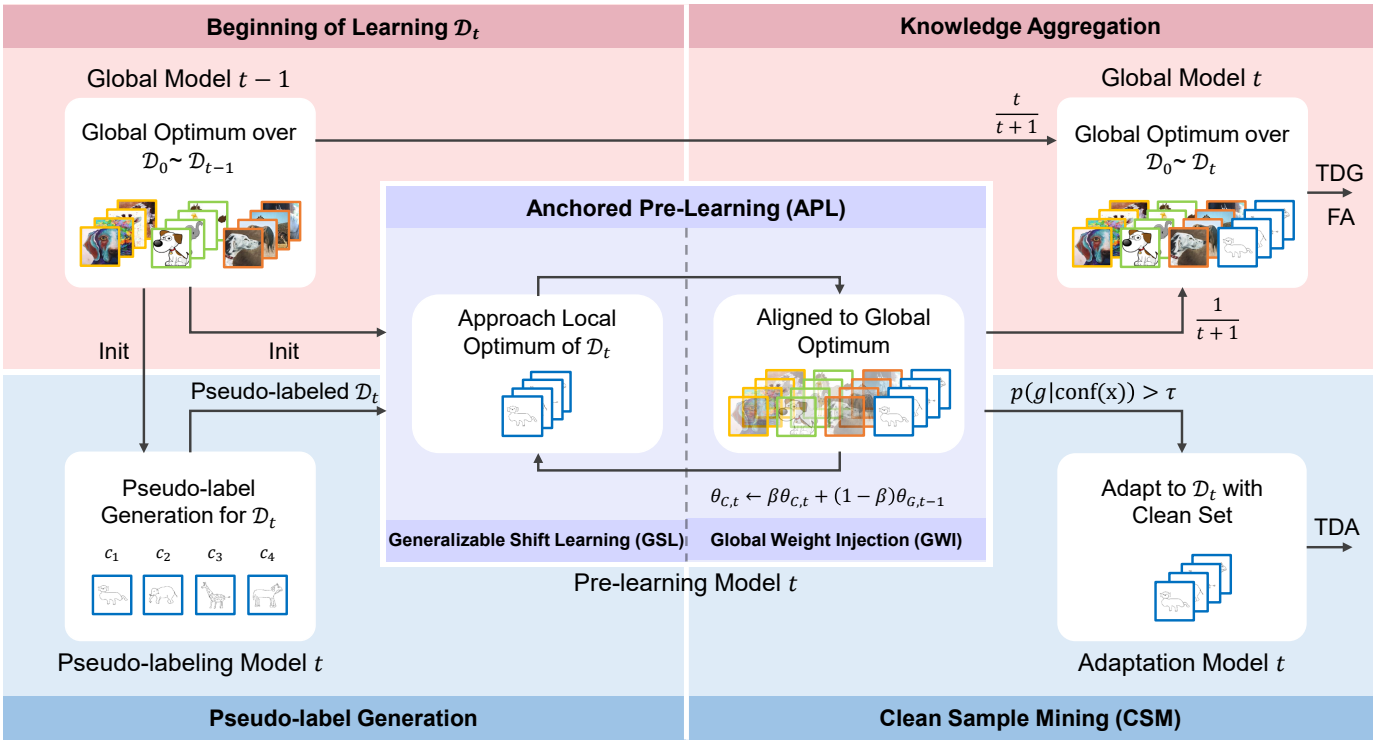

Fig. 1. An overview of how Global-Anchored Learning (GAL) learns the current domain $\mathcal{D}_t$. Based on the previous global model (top left), it first generates pseudo-labels for $\mathcal{D}_t$ (bottom left), and then trains a pre-learning model via APL (center). The knowledge acquired by the pre-learning model is subsequently used to update the global model for TDG and FA (top left), and to construct an adaptation model for $\mathcal{D}_t$-specific TDA (bottom left).

III. Problem Setting

In particular, we consider K-way image classification task. For a vanilla unsupervised CDSL problem, a model $f(x;\theta)$ parameterized by θ is required to sequentially encounter and learn from $T+1$ domains $\mathcal{D}_t\,(t = 0, 1, \ldots, T)$, and achieve reliable performance to map each input image x to its corresponding label y across all the domains. The first domain acts as a labeled source domain $\mathcal{S} = \mathcal{D}_0$ and the others are unlabeled target domains $\mathcal{T} = \{\mathcal{D}_t | t > 0\}$. The model is initialized by $\mathcal{S}$ and then learn from $\mathcal{T}$ in a continual manner. For each time t, we are given a newly arrived domain $\mathcal{D}_t = \{x_t^{(i)}\}_{i=1}^{N_t}$ with N_t unlabeled samples $x_t^{(i)}$, along with a memory buffer $\mathcal{M}_t = \{(x_{t'<t}^{(i)}, \hat{y}_{t'<t}^{(i)})\}_{i=0}^{N_M}$ comprising N_M samples drawn from previously encountered domains, each associated with either the ground-truth label $y_0^{(i)}$ or a pseudo label $\hat{y}_{0<t'<t}^{(i)}$. The goal of unsupervised CDSL is to optimize the model using the current domain $\mathcal{D}_t$ and the memory buffer $\mathcal{M}_t$, enabling it to achieve (i) target domain generalization to unseen domains $\{\mathcal{D}_{t'}\}_{t<t'\leq T}$, (ii) target domain adaptation on the current domain D_t, and (iii) forgetting alleviation, which refers to preserving the performance on past domains $\{\mathcal{D}_{t'}\}_{0\leq t'<t}$.

A. Memory-free CDSL

Unlike existing settings, we tackle a far more challenging yet practical scenario where the model must achieve all three objectives without relying on the assistance of $\mathcal{M}_t$. Although the memory buffer has been widely recognized as a key component for alleviating forgetting and improving generalization [3], we intentionally discard this component to demonstrate the feasibility and effectiveness of a truly memory-free CFDG strategy.

IV. Global-Anchored Learning

To tackle the above memory-free unsupervised CDSL problem, a key challenge lies in the limited exposure of the model to one domain at a time, making it difficult to learn a stable and generalizable decision boundary. In addition, the lack of previous data increases the risk of overfitting to the current domain and leads to catastrophic forgetting, especially under significant distribution shifts.

In order to address these challenges, we propose a novel framework called Global-Anchored Learning (GAL). Instead of retraining a new generalization model from scratch whenever a new domain arrives, GAL incrementally updates a generalizable global model using knowledge from incoming domains, which helps stabilize generalizable decision boundaries and prevent catastrophic forgetting.

As illustrated in Fig 1, GAL completes the learning process for each newly arriving domain in three stages. First, we leverage the previously obtained global model to quickly generate pseudo labels for the current domain via self-supervised learning. Second, we train a pre-learning model with both generalization and adaptation capabilities using the core module of GAL, Anchored Pre-Learning (APL). Third, we utilize the

pre-learning model to identify a clean set of samples by Clean Sample Mining (CSM) for target domain adaptation, while aggregating the knowledge of the pre-learning model into the global model to complete the learning process.

A. Global Model Initialization

Before learning from the target domains ($t = 0$), we consider to initialize a global model $f(\cdot;\theta_{G,0})$ parameterized by $\theta_{G,0}$ as the starting point. We learn it by minimizing the following standard cross-entropy loss,

$$\mathcal{L}_{G,0} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_0} \sum_{k=1}^{K} q_k \log \delta(f(x;\theta_{G,0})), \tag{1}$$

where $\delta_k(a) = \frac{\exp(a_k)}{\sum_i \exp(a_i)}$ denotes the k-th element in the softmax output of a K-dimensional vector a, and q is the one-of-K encoding of y where q_k is 1 for the correct class and 0 for the rest. To ensure both generalization capacity and fairness, we also adopt the same data augmentation strategy RandMix [2] as in prior works.

B. Pseudo-label Generation

Since the current target domain $\mathcal{D}_t$ is unlabeled, it is necessary to generate pseudo-labels for it to facilitate subsequent learning. To this end, we develop a pseudo-labeling model $f(\cdot;\theta_{P,t})$ parameterized by $\theta_{P,t}$. This model is initialized with the previous global model parameters $\theta_{G,t-1}$, thereby leveraging its generalization capability to produce more reliable labels. It first performs unsupervised domain adaptation on $\mathcal{D}_t$ and subsequently generates pseudo-labels based on its predictions.

For unsupervised domain adaptation, we adopt the loss function introduced in the pioneering source-free method SHOT [27], which comprises an information maximization loss and a cross-entropy loss based on self-supervised pseudo-labeling. During training, we freeze the classifier to preserve the generalizable decision boundaries while optimizing only the feature extractor. Notably, we directly employ hard pseudo-labeling via the argmax operation. This design choice aims to ensure methodological simplicity and consistency with later pseudo-labeling stages. The unsupervised domain adaptation process can be written as:

$$\begin{aligned}\mathcal{L}_{P,t} = \mathcal{L}^{im}(\mathcal{D}_t,\, \theta_{P,t}) \\ - \lambda\, \mathbb{E}_{(x,\hat{y})\sim\mathcal{D}_t} \sum_{k=1}^{K} q_k \log \delta(f(x;\theta_{P,t})),\end{aligned} \tag{2}$$

where $\mathcal{L}^{im}$ is the information maximization loss, and $\hat{y}$ is the pseudo label generated after each training epoch, as follows:

$$\hat{y} = \arg\max_{y_k} p(y_k \mid f(x;\theta_{P,t})), \tag{3}$$

where $p(y_k \mid f(x;\theta_{P,t}))$ is the estimated posterior probability of class k given the model output. In addition, Eq. (3) is also the approach used to assign pseudo-labels after adaptation.

C. Anchored Pre-Learning

APL serves as the core module of our framework, playing a crucial role in learning from the current target domain while avoiding overfitting to it. With the aid of pseudo-labels, APL trains a pre-learning model to acquire new knowledge from $\mathcal{D}_t$ while remaining consistent with the existing global knowledge, thereby facilitating subsequent knowledge aggregation and continual learning. This module comprises two subcomponents: GSL and GWI.

a) Generalizable Shift Learning (GSL): Ideally, the global model parameterized by $\theta_{G,t}$ would be optimized using data with ground-truth labels collected from all seen domains. The corresponding training objective can be formulated as:

$$\mathcal{L}_{G,t} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{t'\le t}} \sum_{k=1}^{K} q_k \log \delta(f(x;\theta_{G,t})). \tag{4}$$

For vanilla unsupervised CDSL, the optimization objective defined in Eq. (4) can be approximated by pseudo-labels in conjunction with the replay buffer $\mathcal{M}_t$. The optimization objective can be then reformulated as:

$$\mathcal{L}_{G,t} = -\mathbb{E}_{(x,\hat{y})\sim\mathcal{D}_t\cup\mathcal{M}_t} \sum_{k=1}^{K} q_k \log \delta(f(x;\theta_{G,t})). \tag{5}$$

However, since $\mathcal{M}_t$ is not available, this objective is infeasible in the current setting. When $\theta_{G,t}$ is optimized solely on $\mathcal{D}_t$, it is prone to overfitting to the current domain. Moreover, distillation loss only enforces output similarity and is insufficient to effectively preserve previously acquired knowledge in the presence of significant distribution shifts [3].

To address this challenge, we analyze it from a parameter perspective. When $\theta_{G,t}$ is optimized exclusively on the current domain $\mathcal{D}_t$, it tends to converge to a domain-specific local optimum rather than the desired global optimum, leading to degraded generalization performance and severe catastrophic forgetting.

Fortunately, we observe that this situation closely resembles the local model training in Federated Learning [28], where each local model is trained on distributed clients with distinct data and then aggregated to update a global model. If we treat $\mathcal{D}_0 \sim \mathcal{D}_t$ as data on $t+1$ distributed *time clients*, the update of $\theta_{G,t}$ can be expressed as:

$$\theta_{G,t} = \tfrac{1}{t+1} \sum_{l=0}^{t} \theta_{C,l}, \tag{6}$$

where $\theta_{C,j}$ is the parameter of the domain-specific model of $\mathcal{D}_j$ optimized by the following loss function:

$$\mathcal{L}_{C,l} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_l} \sum_{k=1}^{K} q_k \log \delta(f(x;\theta_{C,l})). \tag{7}$$

It can be observed that, apart from the use of pseudo-labels, Eq. (5) is equivalent to Eq. (7) when the memory buffer $\mathcal{M}_t$ is absent. If we consider $\mathcal{D}_t$ as a new local-specific domain waiting to be learned and $f(\cdot;\theta_{G,t-1})$ as the global model aggregated from domain-specific models of $\mathcal{D}_{t'<t}$, our

learning objective becomes analogous to the scenario of adding a new member to an already completed federated learning process.

Building on this observation, instead of directly training the global model on $\mathcal{D}_t$, we introduce a pre-learning model that serves as a local proxy for the global model to learn from $\mathcal{D}_t$:

$$\mathcal{L}_{C,t} = -\mathbb{E}_{(x,\hat{y})\sim\mathcal{D}_t} \sum_{k=1}^{K} q_k \log \delta(f(x;\theta_{C,t})). \tag{8}$$

For simplicity, we adopt the notation of the domain-specific model in Eq. (7). The pre-learning model is trained exclusively on $\mathcal{D}_t$ to acquire knowledge specific to the current domain. In addition, we employ the same data augmentation strategy as the global model initialization to ensure a certain level of generalization capability for later aggregation. We also employ SelNLPL [29], a method designed to handle noisy labels, to train the pre-learning model.

b) Global Weight Injection (GWI): GWI is introduced to prevent the pre-learning model from overfitting to the current target domain. When trained solely on $\mathcal{D}_t$, the pre-learned model is prone to converge to domain-specific local optimum. Under pronounced distributional shifts, aggregating such model into the global model can undermine global knowledge, a phenomenon known as client drift [30]. Due to communication constraints, federated learning methods typically relies on regularization [31] or aggregation strategies [32] to prevent domain-specific models from drifting excessively away from the global optimum. In our problem, however, since we only need to learn from $\mathcal{D}_t$, communication cost is not a concern. Therefore, we adopt a simpler and more direct approach by increasing the aggregation frequency:

$$\theta_{C,t} \leftarrow \beta\theta_{C,t} + (1-\beta)\theta_{G,t-1}, \tag{9}$$

where β is a hyper-parameter to control the portion of the existing global knowledge.

Unlike task-level parameter interpolation in continual learning, GWI is applied at the end of each epoch. As shown in Eq. (9), a weighted averaging of the local parameters and the existing global parameters is performed after each training epoch. By periodically steering the pre-learned model towards the global optimum, this design promotes the acquisition of current domain knowledge while preserving alignment with the global knowledge.

D. Knowledge Aggregation for Global Model

With the pre-learning model developed by APL, we can aggregate its knowledge of $\mathcal{D}_t$ into the global model without distorting the existing global knowledge:

$$\theta_{G,t} = \tfrac{t}{t+1}\theta_{G,t-1} + \tfrac{1}{t+1}\theta_{C,t}, \tag{10}$$

to approximate Eq. (6). Assuming that $\theta_{G,t-1}$ has reached or is close to the global optimum of $\mathcal{D}_0 \sim \mathcal{D}_{t-1}$, this update ensures that $\theta_{G,t}$ remains aligned with the existing global optimum as it acquires new knowledge over time, thus enhancing generalization and mitigating catastrophic forgetting.

E. Clean Sample Mining

Despite the update of the global model, we consider to develop an distinct adaptation model for each domain, as the purpose of the global optimum is conflict to the local optimum. Since the pre-learning model is trained solely on $\mathcal{D}_t$, it has already achieved reasonably reliable performance on the current domain. However, due to its design objective of enhancing generalization and global optimality, it is not fully adapted to the target domain. To further extract domain-specific knowledge from the pre-learning model for improved adaptation, we draw inspiration from DivideMix [33] and apply a two-component Gaussian Mixture Model (GMM) to identify a clean set of $\mathcal{D}_t$ for domain adaptation.

We perform the separation based on prediction confidence $\text{conf}(x)$ instead of per-sample loss, as the pre-learning model in our case is already fully trained:

$$\text{conf}(x) = \max_{\hat{y}} p(\hat{y} \mid f(x;\theta_{C,t})). \tag{11}$$

We first use the pre-learning model to compute the confidence scores for all samples in $\mathcal{D}_t$ and reassign their pseudo-labels. Then, a two-component Gaussian Mixture Model (GMM) is fitted to these confidence scores, and arrange clean probability for each sample:

$$w(x) = p(g|\text{conf(x)}), \tag{12}$$

where g is the Gaussian component with larger mean (higher confidence). We further distinguish a clean set of $\mathcal{D}_t$ and train the adaptation model:

$$\tilde{\mathcal{D}}_t = \{(x,\hat{y}) \in \mathcal{D}_t \mid w(x) > \tau\}, \tag{13}$$

$$\mathcal{L}_{A,t} = -\mathbb{E}_{(x,\hat{y})\sim\tilde{\mathcal{D}}_t} \sum_{k=1}^{K} q_k \log \delta(f(x;\theta_{A,t})), \tag{14}$$

where τ is a hyper-parameter which control the threshold of the clean probability and $\theta_{A,t}$ denotes the parameters of the adaptation model for $\mathcal{D}_t$.

V. Experiments

A. Setup

For fairness, we follow the experimental settings of existing work [2], [3], including evaluation metrics, dataset and selections of backbones.

a) Evaluation Metrics: Since we evaluate our method on K-way image classification tasks, all metrics represent the accuracy of label predictions. According to the state of each target domain, our evaluation is divided into the following three metrics:

- **TDG**: Target Domain Generalization, which evaluates the performance of the global model parameterized by $\theta_{G,t}$ on unseen target domains $\mathcal{D}_{t+1} \sim \mathcal{D}_T$.
- **TDA**: Target Domain Adaptation, which evaluates the performance of the adaptation model parameterized by $\theta_{A,t}$ on the currently learned target domain $\mathcal{D}_t$.
- **FA**: Forgetting Alleviation, which evaluates the performance of the global model parameterized by $\theta_{G,t}$ on previously learned target domains $\mathcal{D}_0 \sim \mathcal{D}_{t-1}$.

TABLE I

COMPARISONS ON THE PACS, DIGITS-FIVE, AND DOMAINNET DATASETS ARE CONDUCTED ACROSS VARIOUS STATE-OF-THE-ART METHODS UNDER THE TDA, TDG, FA, AND ALL METRICS. THE REPORTED RESULTS ARE AVERAGED OVER 10 DIFFERENT DOMAIN ORDERS WITH CORRESPONDING STANDARD DEVIATION FOR EACH DATASET, AND THE BASELINE RESULTS ARE TAKEN FROM [3]. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD. OUR METHOD CONSISTENTLY OUTPERFORMS ALL BASELINES IN TDG AND TDA, WHILE REMAINING COMPETITIVE IN FA.

Dataset	Metric	SHOT+	SHOT++	Tent	AdaCon	EATA	L2D	PDEN	RaTP	CoDAG	Ours
PACS	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	**72.5** ±7.7
	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	**89.5** ±2.7
	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	87.9 ±3.4
	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	**83.3** ±4.4
Digits-five	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	**77.8** ±5.4
	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	**92.7** ±2.3
	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	**88.5** ±3.7
	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	**86.3** ±3.5
DomainNet	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	**56.8** ±6.5
	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	**75.4** ±4.1
	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	69.8 ±5.5
	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	**67.3** ±5.0

b) Dataset: We consider three widely used datasets in domain generalization tasks. **PACS** [34] contains 7 classes and 4 domains, including Photo (P), Art Painting (A), Cartoon (C) and Sketch (S). **Digits-five** contains 10 classes of numbers, 0 to 9, from 5 domains, including MNIST (MT) [35], MNIST-M (MM) [36], SVHN (SN) [37], SYN-D (SD) [36] and USPS (US) [38]. **DomainNet** [39] contains 10 classes from 6 domains, including Clipart (C), Infograph (I), Painting (P), Quickdraw(Q), Real (R) and Sketch (S). Considering label noise and class imbalance, we follow [2] to construct a subset containing the 10 classes with the largest amount of data for our experiments.

c) Implementation: To ensure fair comparisons, we followed the implementation settings of existing works [2], [40]. For the source domain, 80% of the data is randomly selected for training, with the remaining 20% used for evaluation. For the target domains, all available data is utilized for both training and evaluation. We adopt ResNet-50 [41] as the backbone for PACS and DomainNet, and DTN [27] for the Digits-five dataset. For all datasets, we use the SGD optimizer with a batch size of 64. All results are reported by averaging over 10 randomly sampled domain orders. For each domain order, the performance is further averaged over three independent runs with different random seeds.

B. Results

We first present the comparison results of our method against a series of baselines, followed by ablation studies and hyper-parameter sensitivity analyses.

a) Comparison with baselines: We compare our approach with a series of representative and state-of-the-art methods, covering Source-Free DA (SHOT+ [2], [42] and SHOT++ [42]), Test-Time/Online DA (Tent [43], AdaCon [44] and EATA [45]), Single DG (L2D [20] and PDEN [19]), and replay-based unsupervised CDSL (RaTP [2] and CoDAG [3]), to comprehensively evaluate its performance. As shown in Table I, our method consistently outperforms existing approaches in terms of both target domain generalization and adaptation across all datasets, demonstrating that it effectively drives the model closer to the global optimum while correctly learning the knowledge of the target domains.

In terms of forgetting alleviation, our method surpasses others on Digits-five and remains competitive on PACS and DomainNet, demonstrating that even without a replay buffer, our approach can effectively mitigate forgetting and achieve comparable performance to existing methods.

Moreover, the standard deviation of our method is also competitive compared to existing methods.

b) Ablation Study: We conduct comprehensive ablation studies to validate the effectiveness of the two submodules, GSL and GWI, as well as the downstream module CSM, while source-free adaptation before pseudo-label generation is assumed as a prerequisite following prior CDSL works, with the results summarized in Table II.

First, we investigate whether CSM benefits target domain adaptation. To this end, we remove CSM (Col. 1 & 2), which results in a consistent decline (0.8% on average) in TDA performance. This confirms that extracting target domain knowledge through the GMM-based clean set is both reasonable and beneficial.

Second, we examine the contribution of GWI to mitigating catastrophic forgetting. Removing GWI (Col. 2 & 3) leads to a significant drop in FA performance, highlighting its critical role in alleviating forgetting. The performance on TDG and

TABLE II

ABLATION STUDIES OF GSL, GWI AND CSM ON PACS, DIGITS-FIVE, AND DOMAINNET DATASETS. SINCE APL CONSISTS OF GSL AND GWI, THE LAST LINE IS EQUIVALENT TO ITS ABLATION STUDY.

Module			PACS				Digits-five				DomainNet			
GSL	GWI	CSM	TDG	TDA	FA	ALL	TDG	TDA	FA	ALL	TDG	TDA	FA	ALL
✓	✓	✓	72.5	89.5	87.9	83.3	77.8	92.7	88.5	86.3	56.8	75.4	69.8	67.3
✓	✓	×	72.5	88.9	87.9	83.1	77.8	91.6	88.5	86.0	56.8	74.6	69.8	67.1
✓	×	×	71.4	87.8	83.5	80.9	75.4	91.1	84.7	83.7	54.1	70.2	62.6	62.3
×	×	×	70.8	86.2	82.4	79.8	74.9	90.6	84.1	83.2	53.3	68.1	61.6	61.0

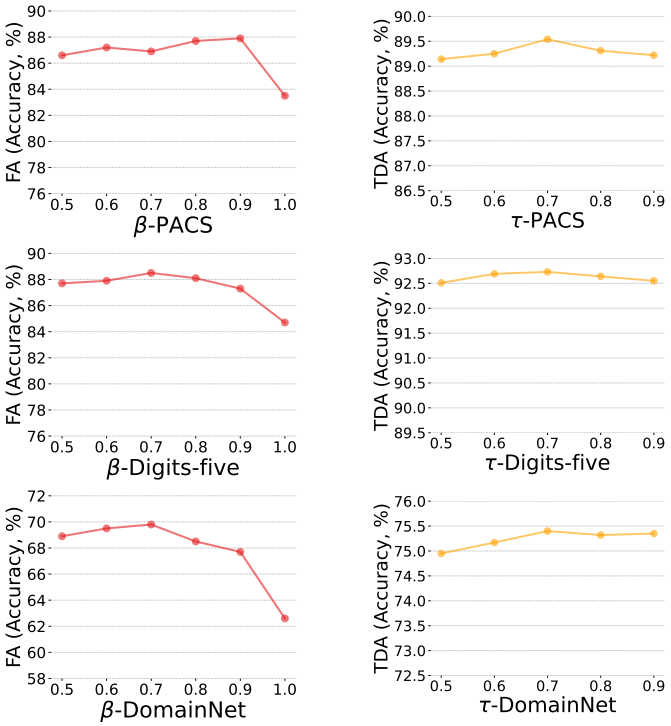

Fig. 2. Sensitivity analysis of β and τ is conducted with metrics most closely related to their respective effects. The results indicate that as long as GWI is activated (i.e., $\beta < 1$), the model can effectively mitigate catastrophic forgetting, and our model is insensitive to τ.

TDA also decreases, indicating that GWI helps the model learn correct semantic knowledge from each target domain.

Finally, we assess the impact of GSL. When GSL is removed (Col. 3 & 4) and the global model is directly aggregated with the pseudo-labeling model, the performance degrades across all metrics. This verifies the effectiveness of GSL in learning from noisy pseudo-labels through data augmentation and denoising.

Note that APL consists of both GSL and GWI, and thus Col. 1 and Col. 3 can be regarded as the ablation study of APL.

c) Hyper-parameter Sensitivity: Our method introduces two hyperparameters: β, controlling the portion of global knowledge injected into the pre-learning model (GWI), and τ, defining the threshold for clean sample selection in CSM. We evaluate their sensitivity across all three datasets.

As shown in Fig. 2, GWI effectively addresses the issue of catastrophic forgetting. Even when a small proportion of $\theta_{G,t-1}$ is incorporated before each training epoch, the pre-learning model remains well consistent with the global model. However, when β is too small (i.e., $\theta_{G,t-1}$ dominates), the model underperforms on FA in practice. We attribute this to the overly dominant global knowledge limiting the ability of the pre-learning model to learn from the current domain, which hinders effective knowledge aggregation and degrades performance in subsequent tasks. We select $\beta = 0.9$ for PACS and $\beta = 0.7$ for both Digits-five and DomainNet to achieve the best performance.

For τ, the results indicate that the model is insensitive to its value. However, when τ is set too low, incorrectly labeled samples may be introduced into the clean set, which leads to a performance degradation (e.g., setting $\tau = 0$, i.e., removing CSM, results in an average TDA drop of 0.8% in Table II). In contrast, an overly large τ yields an excessively small clean set, which increases the risk of overfitting. Based on this trade-off, we set $\tau = 0.7$ to achieve the best overall performance.

VI. CONCLUSION

In this paper, we address a practical Memory-free CDSL setting with the proposed framework named GAL. GAL eliminates the need of preserving data from previously seen domains, thereby offering the feasibility of unsupervised CDSL to scalable, privacy-sensitive and decentralized scenarios. Through theoretical insights, we propose Anchored Pre-Learning (APL), the core of GAL. It leverages two cooperative submodules, Generalizable Shift Learning (GSL) and Global Weight Injection (GWI), to train a pre-learning model that learns from the current domain while staying aligned with the global optimum, thus mitigating catastrophic forgetting. Additionally, Clean Sample Mining (CSM) identifies clean samples to further enhance domain adaptation. Experiments on three benchmark datasets demonstrate that GAL achieves competitive and even state-of-the-art performance.

REFERENCES

[1] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009, vol. 2.

[2] C. Liu, L. Wang, L. Lyu, C. Sun, X. Wang, and Q. Zhu, "Deja vu: Continual model generalization for unseen domains," in *International Conference on Learning Representations*, 2023.

[3] W. Cho, J. Park, and T. Kim, "Complementary domain adaptation and generalization for unsupervised continual domain shift learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 442–11 452.

[4] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[5] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4893–4902.

[6] X. Liu, C. Yoo, F. Xing, H. Oh, G. El Fakhri, J.-W. Kang, J. Woo *et al.*, "Deep unsupervised domain adaptation: A review of recent advances and perspectives," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[7] J. N. Kundu, N. Venkat, R. V. Babu *et al.*, "Universal source-free domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4544–4553.

[8] S. Yang, Y. Wang, J. Van De Weijer, L. Herranz, and S. Jui, "Generalized source-free domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8978–8987.

[9] S. Tang, P. Su, D. Chen, and W. Ouyang, "Gradient regularized contrastive learning for continual domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 2665–2673.

[10] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7201–7211.

[11] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, 2022.

[12] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2100–2110.

[13] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Deep domain-adversarial image generation for domain generalisation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 13 025–13 032.

[14] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[15] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 624–639.

[16] R. Guan, Z. Li, W. Tu, J. Wang, Y. Liu, X. Li, C. Tang, and R. Feng, "Contrastive multiview subspace clustering of hyperspectral images based on graph convolutional networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[17] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[18] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "Metareg: Towards domain generalization using meta-regularization," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[19] L. Li, K. Gao, J. Cao, Z. Huang, Y. Weng, X. Mi, Z. Yu, X. Li, and B. Xia, "Progressive domain expansion network for single domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 224–233.

[20] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, "Learning to diversify for single domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 834–843.

[21] J. Cho, G. Nam, S. Kim, H. Yang, and S. Kwak, "Promptstyler: Prompt-driven style generation for source-free domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 702–15 712.

[22] Y. Tang, Y. Wan, L. Qi, and X. Geng, "Dpstyler: Dynamic promptstyler for source-free domain generalization," *IEEE Transactions on Multimedia*, 2025.

[23] H. Sun, Y. Zhang, L. Xu, S. Jin, P. Luo, C. Qian, W. Liu, and Y. Chen, "Unsupervised continual domain shift learning with multi-prototype modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 10 131–10 141.

[24] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 8, pp. 5362–5383, 2024.

[25] A. Malviya and C. Kumar Maurya, "Hiproibm: unsupervised continual learning through hierarchical prototypical cross-level discrimination along with information bottleneck subnetwork masking," *Applied Intelligence*, vol. 55, no. 6, p. 476, 2025.

[26] S. Mandalika, H. Vardhan, and A. Nambiar, "Replay to remember (r2r): An efficient uncertainty-driven unsupervised continual learning framework using generative replay," *arXiv preprint arXiv:2505.04787*, 2025.

[27] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6028–6039.

[28] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[29] Y. Kim, J. Yim, J. Yun, and J. Kim, "Nlnl: Negative learning for noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 101–110.

[30] Y. Shi, Y. Zhang, Y. Xiao, and L. Niu, "Optimization strategies for client drift in federated learning: A review," *Procedia Computer Science*, vol. 214, pp. 1168–1173, 2022.

[31] L. Gao, H. Fu, L. Li, Y. Chen, M. Xu, and C.-Z. Xu, "Feddc: Federated learning with non-iid data via local drift decoupling and correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 112–10 121.

[32] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7611–7623, 2020.

[33] J. Li, R. Socher, and S. C. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," in *International Conference on Learning Representations*, 2020.

[34] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5542–5550.

[35] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 2002.

[36] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1180–1189.

[37] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng *et al.*, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, vol. 2011. Granada, 2011, p. 7.

[38] J. J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550–554, 2002.

[39] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1406–1415.

[40] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, "Single-source domain expansion network for cross-scene hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 32, pp. 1498–1512, 2023.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[42] J. Liang, D. Hu, Y. Wang, R. He, and J. Feng, "Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8602–8617, 2021.

[43] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," *arXiv preprint arXiv:2006.10726*, 2020.

[44] D. Chen, D. Wang, T. Darrell, and S. Ebrahimi, "Contrastive test-time adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 295–305.

[45] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan, "Efficient test-time model adaptation without forgetting," in *International Conference on Machine Learning*. PMLR, 2022, pp. 16 888–16 905.