frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 03 November 2022
DOI 10.3389/fdata.2022.1052972

OPEN ACCESS

EDITED BY
Shuhan Yuan,
Utah State University, United States

REVIEWED BY
Chao Lan,
University of Oklahoma, United States
Depeng Xu,
University of North Carolina at Charlotte, United States

*CORRESPONDENCE
Jingrui He
jingrui@illinois.edu

SPECIALTY SECTION
This article was submitted to Data Mining and Management, a section of the journal Frontiers in Big Data

RECEIVED 24 September 2022
ACCEPTED 17 October 2022
PUBLISHED 03 November 2022

CITATION
Wu J and He J (2022) Dynamic transfer learning with progressive meta-task scheduler. *Front. Big Data* 5:1052972. doi: 10.3389/fdata.2022.1052972

Dynamic transfer learning with progressive meta-task scheduler

Jun Wu[1] and Jingrui He[1,2]*

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL, United States, [2]School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, IL, United States

Dynamic transfer learning refers to the knowledge transfer from a static source task with adequate label information to a dynamic target task with little or no label information. However, most existing theoretical studies and practical algorithms of dynamic transfer learning assume that the target task is continuously evolving over time. This strong assumption is often violated in real world applications, e.g., the target distribution is suddenly changing at some time stamp. To solve this problem, in this paper, we propose a novel meta-learning framework L2S based on a progressive meta-task scheduler for dynamic transfer learning. The crucial idea of L2S is to incrementally learn to schedule the meta-pairs of tasks and then learn the optimal model initialization from those meta-pairs of tasks for fast adaptation to the newest target task. The effectiveness of our L2S framework is verified both theoretically and empirically.

KEYWORDS

transfer learning, distribution shift, dynamic environment, meta-learning, task scheduler, image classification

1. Introduction

Transfer learning (Pan and Yang, 2009; Tripuraneni et al., 2020) improves the generalization performance of a learning algorithm on the target task, by leveraging the knowledge from a relevant source task. It has been studied (Ben-David et al., 2010; Long et al., 2015; Ganin et al., 2016; Zhang et al., 2019) that the knowledge transferability across tasks can be theoretically guaranteed under mild conditions, e.g., source and target tasks share the same labeling function. One assumption behind those works is that source and target tasks are sampled from a stationary task distribution. More recently, it is observed that in the context of transfer learning, the tasks might be sampled from a non-stationary task distribution, i.e., the learning task might be evolving over time in real scenarios. It can be formulated as a dynamic transfer learning problem from a static source task[1] with adequate label information to a dynamic target task with little or no label information (see Figure 1).

1 It can also be generalized to the scenarios (Wu and He, 2022b) where the knowledge is transferred from a dynamic source task to a dynamic target task.

Most existing works (Hoffman et al., 2014; Bobu et al., 2018; Kumar et al., 2020; Wang H. et al., 2020; Wu and He, 2020, 2022b) on dynamic transfer learning assume that the target task is continuously changing over time. This assumption allows deriving the generalization error bound of dynamic transfer learning using the distribution shift at any consecutive time stamps. Nevertheless, we show that these error bounds are not tight when the task distribution changes suddenly at some time stamp. Therefore, previous works can be hardly applied to real scenarios where the task distribution might not always be evolving continuously. This sudden distribution shift can be induced by some unexpected issues, e.g., adversarial attacks (Wu and He, 2021), system failures (Lu et al., 2018), etc.

To solve this problem, we derive the generalization error bound of dynamic transfer learning in terms of adaptively scheduled meta-pairs of tasks. Moreover, it is observed that this result is closely related to the existing error bounds (Wang et al., 2022; Wu and He, 2022b). It is found that previous works showed the error bounds in terms of the distribution shift at any consecutive time stamps. In contrast, we consider all the meta-pairs of tasks, e.g., a pair of tasks transferring the knowledge from an old time stamp to a new time stamp. As a result, our error bound can be tight even when the task distribution is suddenly shifted at some time stamp. Then, by minimizing the error bound, we propose a novel meta-learning framework L2S based on a progressive meta-task scheduler for dynamic transfer learning. In this framework, we automatically learn the sampling probability for meta-pairs of tasks based on task relatedness. The effectiveness of L2S framework is then verified on a variety of dynamic transfer learning tasks. The major contributions of this paper are summarized as follows.

- We consider a relaxed assumption of dynamic transfer learning, i.e., the target task distribution might change suddenly at some time stamp when it is evolving over time. The generalization error bounds of dynamic transfer learning can then be derived with this relaxed assumption.
- We propose a novel meta-learning framework L2S based on a progressive meta-task scheduler for dynamic transfer learning. Different from recent work (Wu and He, 2022b), L2S learns to schedule the meta-pairs of tasks based on task relatedness.
- Experiments on various data sets demonstrate the effectiveness of our L2S framework over state-of-the-art baselines.

The rest of the paper is organized as follows. We review the related work in Section 2. The problem of dynamic transfer learning is defined in Section 3. In Section 4, we derive the error bounds of dynamic transfer learning, followed by the proposed L2S framework in Section 5. The empirical analysis on L2S is provided in Section 6. Finally, we conclude the paper in Section 7.

2. Related work

In this section, we briefly introduce the related work on dynamic transfer learning and meta-learning.

2.1. Dynamic transfer learning

Dynamic transfer learning (Hoffman et al., 2014; Bitarafan et al., 2016; Mancini et al., 2019) refers to the knowledge transfer from a static source task to a dynamic target task. Compared to standard transfer learning on the static source and target tasks (Pan and Yang, 2009; Zhou et al., 2017, 2019a,b; Tripuraneni et al., 2020; Wu and He, 2021), dynamic transfer learning is a more challenging but realistic problem setting due to its time evolving task relatedness. More recently, various dynamic transfer learning frameworks are built from the following aspects: self-training (Kumar et al., 2020; Chen and Chao, 2021; Wang et al., 2022), incremental distribution alignment (Bobu et al., 2018; Wulfmeier et al., 2018; Wang H. et al., 2020; Wu and He, 2020, 2022a), meta-learning (Liu et al., 2020; Wu and He, 2022b), contrastive learning (Tang et al., 2021; Taufique et al., 2022), etc. Specifically, most existing works assume that the task distribution is continuously evolving over time. Very little effort has been devoted to studying dynamic transfer learning when this assumption is violated in real scenarios. Compared to previous works (Liu et al., 2020; Wang et al., 2022; Wu and He, 2022b), in this paper, we focus on a more realistic dynamic transfer learning with a relaxed assumption that the task distribution could be suddenly changed at some time stamp.

2.2. Meta-learning

Meta-learning (Hospedales et al., 2021) leverages the knowledge from a set of prior meta-training tasks for fast adaptation to new tasks. In the context of few-shot classification, meta-learning aims to find the optimal model initialization (Finn et al., 2017, 2018; Wang L. et al., 2020; Yao et al., 2021) from previously seen tasks such that this model can be fine-tuned on a new task by performing a few gradient steps. It assumes that all the tasks follow a stationary task distribution. More recently, this meta-learning paradigm has been extended into the online learning setting where a sequence of tasks is sampled from non-stationary task distributions (Finn et al., 2019; Acar et al., 2021). Following previous work (Wu and He, 2022b), we formulate dynamic transfer learning as a meta-learning problem, which aims to learn the optimal model initialization for knowledge transfer across any meta-pair of tasks. In contrast to Wu and He (2022b) where the meta-pairs of tasks are simply constructed from tasks at consecutive time stamps, we propose to learn the sampling probability for meta-pairs of tasks based on the task relatedness during model

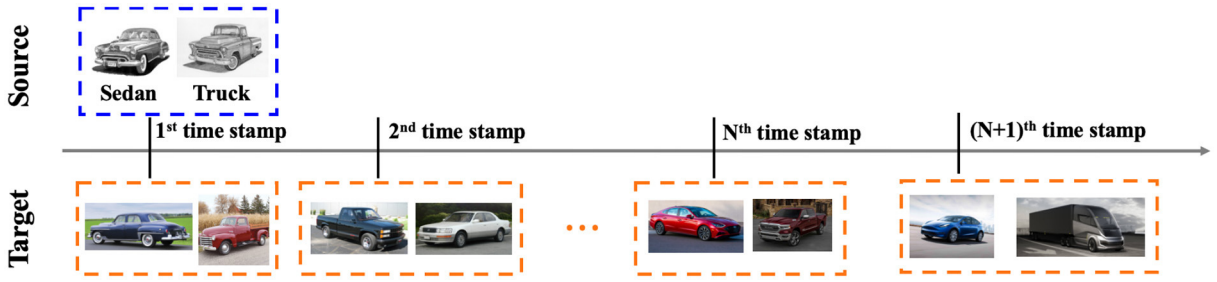

FIGURE 1
Illustration of dynamic transfer learning from a static source task (e.g., sketch image classification with fully labeled examples) to a dynamic target task (e.g., real-world image classification with only unlabeled examples).

training. This can help our meta-learning framework avoid the negative transfer induced by the meta-pairs of tasks sampled from suddenly shifted task distribution.

3. Preliminaries

In this section, we present the notation and formal problem definition of dynamic transfer learning.

3.1. Notation

Let $\mathcal{X}$ and $\mathcal{Y}$ be the input feature space and output label space respectively. We consider the dynamic transfer learning problem (Hoffman et al., 2014; Bobu et al., 2018) with a static source task $\mathcal{D}^s$ and a dynamic target task $\{\mathcal{D}_j^t\}_{j=1}^N$ with time stamp j. In this case, we assume that there are m^s labeled training examples $D^s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{m^s}$ in the source task. Let m_j^t be the number of unlabeled training examples $D_j^t = \{\mathbf{x}_{ij}^t\}_{i=1}^{m_j^t}$ in the j^{th} target task. Let $\mathcal{H}$ be the hypothesis class on $\mathcal{X}$ where a hypothesis is a function $h : \mathcal{X} \to \mathcal{Y}$. $\mathcal{L}(\cdot, \cdot)$ is the loss function such that $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The expected classification error on the source task $\mathcal{D}^s$ is defined as $\epsilon^s(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}^s}[\mathcal{L}(h(\mathbf{x}), y)]$ for any $h \in \mathcal{H}$, and its empirical estimate is given by $\hat{\epsilon}^s(h) = \frac{1}{m^s}\sum_{i=1}^{m^s}\mathcal{L}(h(\mathbf{x}_i), y_i)$. The expected error $\epsilon_j^t(h)$ and empirical error $\hat{\epsilon}_j^t(h)$ of the target task at the j^{th} time stamp can also be defined similarly.

3.2. Problem definition

Following previous works (Hoffman et al., 2014; Bitarafan et al., 2016; Bobu et al., 2018), we formally define the problem of dynamic transfer learning as follows.

Definition 3.1. *(Dynamic Transfer Learning) Given a labeled static source task $\mathcal{D}^s$ and an unlabeled dynamic target task $\{\mathcal{D}_j^t\}_{j=1}^N$, dynamic transfer learning aims to learn the prediction function for the newest target task $\mathcal{D}_{N+1}^t$ by leveraging the knowledge from historical source and target tasks.*

The key challenge of dynamic transfer learning is the time evolving task relatedness between source and target tasks. Recent works (Liu et al., 2020; Wang et al., 2022; Wu and He, 2022b) showed the generalization error bounds by assuming that the data distribution of the target task is continuously changing over time. Intuitively, in this case, the expected error bound on the newest target task is bounded in terms of the largest distribution gap [e.g., $\max_{0\le j\le N} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_j^t, \mathcal{D}_{j+1}^t)$] across time stamps. As a result, these generalization error bounds are not tight when the task distribution is significantly shifted at some time stamp. As shown in Figure 2, the task distribution is shifted smoothly from time stamp 1 to time stamp 2. However, it changes sharply from time stamp 2 to time stamp 3. In real scenarios, this sharp distribution shift might be induced by some unexpected issues, e.g., adversarial manipulation (Wu and He, 2021). This thus motivates us to study dynamic transfer learning with a much more relaxed assumption that the task distribution could be suddenly shifted at some time stamp.

4. Theoretical analysis

In this section, we provide the theoretical analysis for dynamic transfer learning.

4.1. Generalization error bound

We derive the generalization error bound of dynamic transfer learning as follows. Following Ben-David et al. (2010) and Liu et al. (2020), we use $\mathcal{H}$-divergence to measure the distribution shift across tasks and Vapnik-Chervonenkis (VC) dimension to measure the complexity of a class of functions $\mathcal{H}$. Without loss of generality, we would like to consider a binary classification problem (i.e., $\mathcal{Y} = \{0, 1\}$) with the loss function

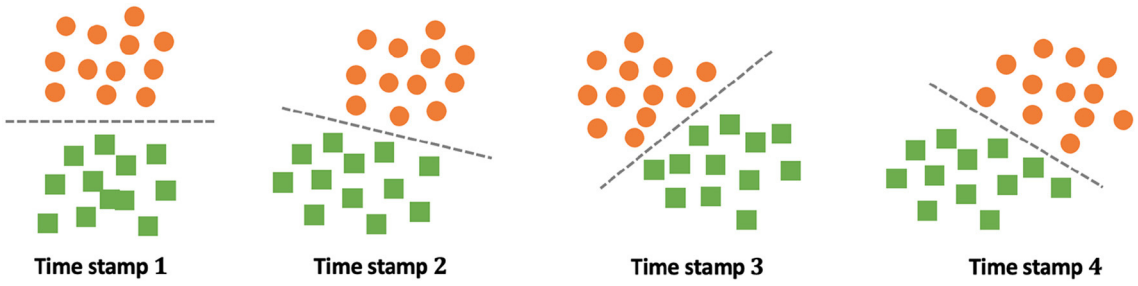

FIGURE 2
Challenges of dynamic transfer learning where the task distribution is suddenly changed at time stamp 3. Here orange circle and green square denote data points from two classes, and the dashed line indicates the optimal decision boundary at different time stamps.

$\mathcal{L}(\hat{y}, y) = |\hat{y} - y|$. The following theorem showed that the expected error of the newest target task $\mathcal{D}^t_{N+1}$ can be bounded in terms of the historical source and target knowledge.

Theorem 4.1. *(Generalization Error Bound) Let $\mathcal{H}$ be a hypothesis space of VC dimension d. If there are m labeled source examples i.i.d. drawn from $\mathcal{D}^s$ (denoted as $\mathcal{D}^t_0$ as well) and m unlabeled target examples i.i.d. drawn from $\mathcal{D}^t_j$ for each time stamp $j = 1, \cdots, N+1$[2], then for any $\delta > 0$ and $h \in \mathcal{H}$, with probability at least $1 - \delta$, the expected error of the newest target task $\mathcal{D}^t_{N+1}$ can be bounded as follows.*

$$\epsilon^t_{N+1}(h) \leq \sum_{i=0}^{N} \sum_{j=i+1}^{N+1} w_{ij} \left(\hat{\epsilon}^t_i(h) + \eta_{ij} \cdot \hat{d}_{\mathcal{H}\Delta\mathcal{H}} \left(\mathcal{D}^t_i, \mathcal{D}^t_j \right) \right) + \mathcal{O}\left(\lambda + \sqrt{\frac{d \log(2m) + \log(2/\delta) + \sum_{i=0}^{N} \sum_{j=i+1}^{N+1} w_{ij}^2 \log(1/\delta)}{2m}} \right)$$

where $\sum_{i=0}^{N} \sum_{j=i+1}^{N+1} w_{ij} = 1$, and $w_{ij} \geq 0$ if $i < j$, $w_{ij} = 0$ otherwise. $\eta_{ij} = \frac{1}{2}$ if $1 \leq j \leq N$ and $i < j$, and $\eta_{ij} = \frac{1}{2}\left(1 + \frac{\sum_{k=0}^{i-1} w_{ki}}{w_{ij}}\right)$ if $j = N+1$ and $i < j$, $\eta_{ij} = 0$ otherwise. Here λ denotes the combined error of the ideal hypothesis over all the tasks, i.e., $\lambda = \min_{h \in \mathcal{H}} \sum_{i=0}^{N+1} \epsilon^t_i(h)$, and $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\cdot, \cdot)$ denotes the empirical estimate of $\mathcal{H}$-divergence over finite examples.

Note that this error bound holds with other existing distribution discrepancy measures (see Corollary 4.3), though we consider $\mathcal{H}$-divergence (Ben-David et al., 2010) in Theorem 4.1. Furthermore, we show the generalization error bound of dynamic transfer learning from the perspective of meta-learning. That is, instead of sharing the hypothesis $h \in \mathcal{H}$ for all the tasks, we learn a common initialized model $\bar{h} \in \mathcal{H}$ across tasks. Then the task-specific model h_i via one-step gradient update for the target at the i^{th} time stamp, i.e., $\theta_i = \bar{\theta} - \beta \nabla_\theta \mathcal{L}^{meta}$, where $\theta_i, \bar{\theta}$ denotes the parameters of $h_i, \bar{h}$ respectively and $\mathcal{L}^{meta}$ is the meta-learning loss for updating the task-specific model parameters. If we let $\mathcal{L}^{meta} = \hat{\epsilon}^t_i(\bar{h}) = \frac{1}{m} \sum_{k=1}^{m} \mathcal{L}[\bar{h}(\mathbf{x}_{ki}), y_{ki}]$, the following theorem provides the generalization error bound based on meta-learning.

Theorem 4.2. *(Meta-Learning Generalization Error Bound) Let $\mathcal{H}$ be a hypothesis space of VC dimension d. If there are m labeled source examples i.i.d. drawn from $\mathcal{D}^s$ (denoted as $\mathcal{D}^t_0$ as well) and m unlabeled target examples i.i.d. drawn from $\mathcal{D}^t_j$ for each time stamp $j = 1, \cdots, N+1$, then for any $\delta > 0$ and a proper inner learning rate β, with probability at least $1 - \delta$, the expected error of the newest target task $\mathcal{D}^t_{N+1}$ can be bounded in the following.*

$$\epsilon^t_{N+1}(h_{N+1}) \leq \sum_{i=0}^{N} \sum_{j=i+1}^{N+1} w_{ij} \left(\hat{\epsilon}^t_i(h_i) + \eta_{ij} \cdot \hat{d}_{\mathcal{H}\Delta\mathcal{H}} \left(\mathcal{D}^t_i, \mathcal{D}^t_j \right) \right) + \mathcal{O}\left(\sum_{i=0}^{N} \left(\frac{1}{m} \sum_{k=1}^{m} \left|\left| \nabla_\theta \bar{h}(\mathbf{x}_{ki}) \right|\right| \right)^2 + \lambda + \sqrt{\frac{d \log(2m) + \log(2/\delta) + \sum_{i=0}^{N} \sum_{j=i+1}^{N+1} w_{ij}^2 \log(1/\delta)}{m}} \right)$$

where $\sum_{i=0}^{N} \sum_{j=i+1}^{N+1} w_{ij} = 1$, and $w_{ij} \geq 0$ if $i < j$, $w_{ij} = 0$ otherwise. $\eta_{ij} = \frac{1}{2}$ if $1 \leq j \leq N$ and $i < j$, and $\eta_{ij} = \frac{1}{2}\left(1 + \frac{\sum_{k=0}^{i-1} w_{ki}}{w_{ij}}\right)$ if $j = N+1$ and $i < j$, $\eta_{ij} = 0$ otherwise. Here λ denotes the combined error of the ideal hypothesis over all the tasks, i.e., $\lambda = \min_{h \in \mathcal{H}} \sum_{i=0}^{N+1} \epsilon^t_i(h)$, and $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\cdot, \cdot)$ denotes the empirical estimate of $\mathcal{H}$-divergence over finite examples.

We observe from Theorem 4.2 that the parameter w_{ij} plays an important role in the generalization error bound of dynamic transfer learning. Intuitively, it is more likely to assign higher value w_{ij} for the easy meta-pair of tasks $\mathcal{D}_i \rightarrow \mathcal{D}_j$

2 Here we assume that it generates the same number of examples at every time stamp, i.e., $m^s = m^t_1 = \cdots = m^t_{N+1} = m$, but the theoretical results can also be generalized into the scenarios with different number of samples in source and target tasks.

with stronger class discrimination over $\mathcal{D}_i$ [i.e., smaller $\hat{\epsilon}_i^t(h_i)$] and smaller distribution shift between $\mathcal{D}_i$ and $\mathcal{D}_j$ [i.e., smaller $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_i^t, \mathcal{D}_j^t)$].

4.2. Connection to existing bounds

The following corollary shows that the error bound in Theorem 4.1 can be generalized by considering various domain discrepancy measures.

Corollary 4.3. *With the same assumptions in Theorem 4.1, for any $\delta > 0$ and $h \in \mathcal{H}$, there exist $w_{ij} \geq 0$ and $\eta_{ij} \geq 0$, with probability at least $1 - \delta$, the expected error of the newest target task $\mathcal{D}_{N+1}^t$ can be bounded in the following.*

$$\epsilon_{N+1}^t(h) \leq \sum_{i=0}^{N} \sum_{j=i+1}^{N+1} w_{ij} \left(\hat{\epsilon}_i^t(h) + \eta_{ij} \cdot \hat{d}\left(\mathcal{D}_i^t, \mathcal{D}_j^t \right) \right) + \Omega \quad (1)$$

where $\hat{d}(\cdot, \cdot)$ can be instantiated with existing distribution discrepancy measures, including discrepancy distance (Mansour et al., 2009), maximum mean discrepancy (Long et al., 2015), Wasserstein distance (Shen et al., 2018), f-divergence (Acuna et al., 2021), etc. Here Ω denotes the corresponding sample complexity when the distribution discrepancy measure is selected.

Corollary 4.3 shows the flexibility in generalizing existing static transfer learning theories (Mansour et al., 2009; Ben-David et al., 2010; Ghifary et al., 2016; Shen et al., 2018; Zhang et al., 2019; Acuna et al., 2021) into the dynamic transfer learning setting. Moreover, it is observed that Corollary 4.3 is closely related to the existing generalization error bounds (Wang et al., 2022; Wu and He, 2022b) of dynamic transfer learning, under different parameters w_{ij} and η_{ij}.

- When w_{ij} and η_{ij} are given by

$$w_{ij} = \begin{cases} \frac{1}{N+1}, & \text{if } i = 0 \\ \frac{\tau}{N+1}, & \text{if } 1 \leq i \leq N \text{ and } i + 1 = j \\ 0, & \text{otherwise} \end{cases}$$

$$\eta_{ij} = \begin{cases} \rho\sqrt{R^2+1}(N+1), & \text{if } i = 0 \text{ and } j = 1 \\ \rho\sqrt{R^2+1}(N+1)/\tau, & \text{if } 1 \leq i \leq N \text{ and } i + 1 = j \\ 0, & \text{otherwise} \end{cases}$$

where $\tau \in \mathbb{R}$. Then, when $\tau \to 0$, Corollary 4.3 recovers the generalization error bound (Wang et al., 2022).

$$\epsilon_{N+1}^t(h_{N+1}) \leq \epsilon^s(h_0) + \rho\sqrt{R^2+1} \sum_{i=1}^{N+1} d_{W_p}\left(\mathcal{D}_{i-1}^t, \mathcal{D}_i^t \right) + \mathcal{O}\left(N\sqrt{\frac{\log(1/\delta)}{m}} + \frac{N}{\sqrt{m}} + \frac{1}{\sqrt{mN}} + \sqrt{\frac{\log(mN)^{3L-2}}{mN}} + \sqrt{\frac{\log(1/\delta)}{mN}} \right)$$

where $\mathcal{H}$ is the hypothesis class of R-Lipschitz L-layer fully-connected neural networks with 1-Lipschitz activation function.

- When w_{ij} and η_{ij} are given by

$$w_{ij} = \begin{cases} \frac{1}{N+1}, & \text{if } i + 1 = j \\ 0, & \text{otherwise} \end{cases} \qquad \eta_{ij} = \begin{cases} 1, & \text{if } i + 1 = j \\ 0, & \text{otherwise} \end{cases}$$

Then, Corollary 4.3 recovers the generalization error bound (Wu and He, 2022b).

$$\epsilon_{N+1}^t(h) \leq \sum_{i=1}^{N+1} \frac{1}{N+1} \left(\hat{\epsilon}_{i-1}^t(h) + \hat{d}_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{D}_{i-1}^t, \mathcal{D}_i^t \right) \right) + \Omega_L \quad (2)$$

where Ω_L is a Rademacher complexity term.

Compared to existing theoretical results (Wang et al., 2022; Wu and He, 2022b), with appropriate w_{ij}, our generalization error bound in Corollary 4.3 is much more tighter when there exists some time stamp i such that $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{D}_{i-1}^t, \mathcal{D}_i^t \right)$ is large. It thus motivates us to develop a progressive meta-task scheduler in the meta-learning framework for dynamic transfer learning. The crucial idea is to automatically learn the values w_{ij}, based on the intuition that assigning large value w_{ij} on easy meta-pair of tasks $\mathcal{D}_i \to \mathcal{D}_j$ would make our error bound much tighter.

5. Methodology

Following Wu and He (2022b), we propose a meta-learning framework named L2S for dynamic transfer learning by empirically minimizing the error bound in Theorem 4.2. Instead of uniformly sampling the meta-pairs of tasks in the consecutive time stamps (Wu and He, 2022b), in this paper, we learn a progressive meta-task scheduler for automatically formulating the meta-pairs of tasks from the dynamic target task.

The overall objective function of L2S for learning the prediction function of $\mathcal{D}_{N+1}^t$ on the $(N+1)^{\text{th}}$ time stamp is

given as follows.

$$\begin{aligned}&\min_{\theta}\min_{\mathbf{w}}\mathcal{J}(\theta,\mathbf{w})\\&=\sum_{i=0}^{N}\sum_{j=i+1}^{N+1}w_{ij}\left(\hat{\epsilon}_i^t(M_{ij}(\theta))+\eta\cdot\hat{d}_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{D}_i^t,\mathcal{D}_j^t;M_{ij}(\theta)\right)\right)\\&\text{s.t.}\quad\sum_{i=0}^{N}\sum_{j=i+1}^{N+1}w_{ij}=1\\&\text{s.t.}\quad M_{ij}(\theta)=\theta-\beta\nabla_\theta\mathcal{L}^{meta}(\mathcal{D}_i^t,\mathcal{D}_j^t)\end{aligned}\tag{3}$$

where θ is the trainable parameters and $\mathcal{L}^{meta}(\mathcal{D}_i^t,\mathcal{D}_j^t)$ is the meta-training loss. $\eta\geq 0$ is a hyper-parameter to balance the classification error and discrepancy minimization.

The proposed L2S framework has three crucial components: meta-pairs of tasks, meta-training, and meta-testing. The overall training procedures of L2S are illustrated in Algorithm 1.

- **Meta-Pairs of Tasks:** Following the theoretical results in Section 4.1, we formulate the candidate meta-pairs of tasks from any two different time stamps $(\mathcal{D}_i^t,\mathcal{D}_j^t)$ $(i<j)$. It can be considered as a simple knowledge transfer from $\mathcal{D}_i^t$ to $\mathcal{D}_j^t$. Here we simply denote the source task $\mathcal{D}^s$ as $\mathcal{D}_0^t$. Since we focus on learning the prediction function on the target task at a new time stamp, we consider the knowledge transfer from an old time stamp i to a new time stamp j, i.e., $i<j$. Note that as suggested in Theorem 4.2, those candidate meta-pairs of tasks might not have equal sampling probability for meta-training. Therefore, we propose a progressive meta-pair scheduler to incrementally learn the sampling probability of every candidate meta-pair of tasks.

  As shown in Theorem 4.2, the sampling probability w_{ij} is strongly related to the classification error on $\mathcal{D}_i^t$ and the empirical distribution discrepancy between $\mathcal{D}_i^t$ and $\mathcal{D}_j^t$. However, we have only unlabeled training examples for the target task. It is intractable to accurately estimate the classification error on $\mathcal{D}_i^t$ $(i=1,2,\cdots)$ for the target task. One solution is that we can incrementally estimate the pseudo-labels of unlabeled target examples, and then obtain the classification error using these pseudo-labels. But it will be largely affected by the quality of the pseudo-labels. Instead, in this paper, we simply learn the sampling probability using the empirical distribution discrepancy between $\mathcal{D}_i^t$ and $\mathcal{D}_j^t$ because this distribution discrepancy involves only the unlabeled examples. That is, the sampling probability w_{ij} is learned as follows.

$$w_{ij}=\frac{\exp\left(1/\hat{d}_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{D}_i^t,\mathcal{D}_j^t\right)\right)}{\Gamma}\tag{4}$$

  where Γ is a normalization term. it indicates that the meta-pair of tasks with a smaller distribution discrepancy has a larger probability of being sampled for meta-training. Intuitively, the smaller distribution discrepancy guarantees the knowledge transfer across tasks (Ganin et al., 2016; Zhang et al., 2019). Therefore, we can sample a set of meta-pairs of tasks $\mathcal{S}$ based on the sampling probability for meta-training.
- **Meta-Training:** Following Wu and He (2022b), the meta-training over meta-pairs of tasks is given as follows. Let $\zeta_{ij}(\theta)=\hat{\epsilon}_i^t(M_{ij}(\theta))+\eta\cdot\hat{d}_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{D}_i^t,\mathcal{D}_j^t;M_{ij}(\theta)\right)$ be the loss function over the validation set on a meta-pair of tasks. Then the model initialization θ can be learned by

$$\begin{aligned}\theta&\leftarrow\arg\min_{\theta}\sum_{(i,j)\in\mathcal{S}}\zeta_{ij}(\theta)\\M_{ij}(\theta)&\leftarrow\theta-\beta\nabla_\theta\mathcal{L}^{meta}(\mathcal{D}_i^t,\mathcal{D}_j^t)\end{aligned}\tag{5}$$

  where $M_{ij}:\theta\rightarrow\theta_{ij}$ is a function which maps the model initialization θ into the optimal task-specific parameter θ_{ij}. Similar to the model-agnostic meta-learning (MAML) (Finn et al., 2017), $M_{ij}(\theta)$ can be instantiated by one or a few gradient descent updates in practice. In this case, the meta-training loss is given by $\mathcal{L}^{meta}(\mathcal{D}_i^t,\mathcal{D}_j^t)=\hat{\epsilon}_i^t(M_{ij}(\theta))+\eta\cdot\hat{d}_{\mathcal{H}\Delta\mathcal{H}}\left(\mathcal{D}_i^t,\mathcal{D}_j^t;M_{ij}(\theta)\right)$ over the training set.

  As illustrated in Algorithm 1, the predictive function is incrementally learned for the target task at every historical time stamp, and then the pseudo-labels of unlabeled target examples can be inferred.
- **Meta-Testing:** The optimal parameters θ_{N+1} on the newest target task $\mathcal{D}_{N+1}^t$ could be learned by fine-tuning the optimal model initialization θ on a selective meta-pair of tasks $(\mathcal{D}_k^t,\mathcal{D}_{N+1}^t)$.

$$\theta_{N+1}=M_{k(N+1)}(\theta)\leftarrow\theta-\beta\nabla_\theta\mathcal{L}^{meta}(\mathcal{D}_k^t,\mathcal{D}_{N+1}^t)\tag{6}$$

  where θ is the optimized model initialization learned in the meta-training phase. Here we choose the meta-pair of tasks $(\mathcal{D}_k^t,\mathcal{D}_{N+1}^t)$ by estimating the sampling probability $w_{k(N+1)}$ $(k=0,1,\cdots,N)$ and choosing k with the largest value $w_{k(N+1)}$.

6. Experiments

In this section, we provide the empirical analysis of L2S framework on various data sets.

6.1. Experimental setup

We used the following publicly available image data sets:

- Rotating MNIST (Kumar et al., 2020): The original MNIST (LeCun et al., 1998) is a digital image data set with 60,000 images from 10 categories. Rotating MNIST is a semi-synthetic version of MNIST where each image is rotated by a degree. Following Bobu et al. (2018) and Kumar et al. (2020), we rotate each image by an angle

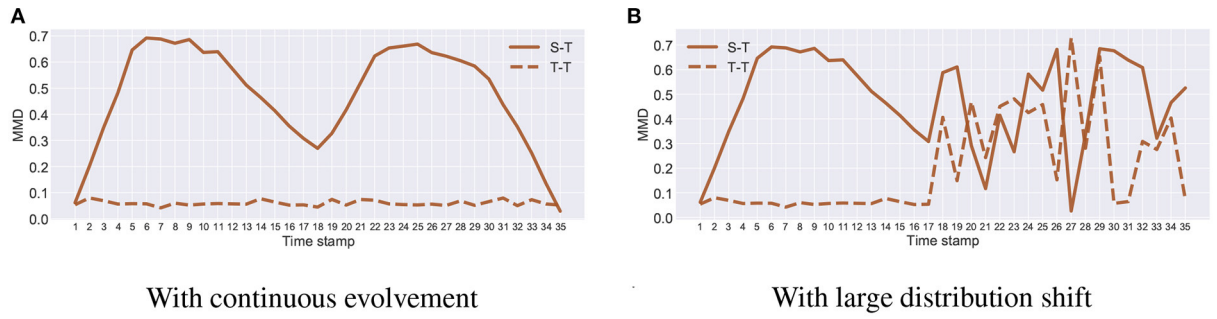

FIGURE 3
Rotating MNIST with **(A)** continuous evolvement and **(B)** large distribution shift.

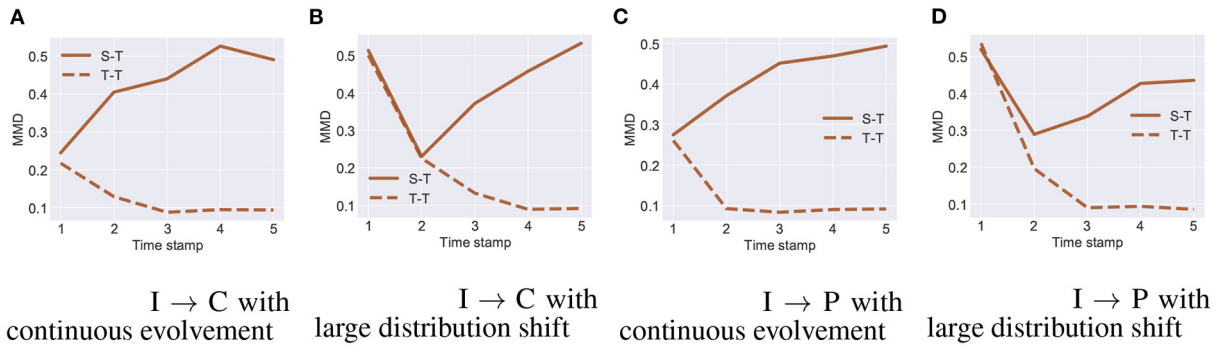

FIGURE 4
I → C on ImageCLEF-DA with **(A)** continuous evolvement, **(B)** large distribution shift. I → P on ImageCLEF-DA with **(C)** continuous evolvement, **(D)** large distribution shift.

TABLE 1 Results of dynamic transfer learning on Rotating MNIST.

Methods	With continuous evolvement		With large distribution shift	
	Acc	H-Acc	Acc	H-Acc
SourceOnly	1.0000	0.4393	0.3437	0.4393
DAN (Long et al., 2015)	1.0000	0.4518	0.5625	0.4830
DANN (Ganin et al., 2016)	1.0000	0.3884	0.3750	0.4000
MDD (Zhang et al., 2019)	1.0000	0.4250	0.4063	0.4482
CUA (Bobu et al., 2018)	0.9375	0.9277	0.4375	0.8259
GST (Kumar et al., 2020)	0.0625	0.1062	0.1250	0.2259
L2E (Wu and He, 2022b)	0.9688	0.9795	0.6250	0.7179
L2S	**1.0000**	**0.9991**	**0.9687**	**0.9116**

The best results are indicated in bold.

for generating the time-evolving classification task. More specifically, for the source task, we randomly choose 32 images and then rotate them by an angle between 0 and 10 degrees. All the images in the source task are associated with class labels. For the time-evolving target task, we randomly choose 32 images at every time stamp j ($j = 1, \cdots, 35$) and rotate them by an angle between $10 \cdot j$ and $10 \cdot (j+1)$ degrees. It can be seen that in this case, the data distribution of the target task is continuously evolving over time. Therefore, we denote the aforementioned Rotating MNIST as a data set "with continuous evolvement." In contrast, we consider the dynamic transfer learning scenarios "with large distribution shift," where the samples at the last 18 time stamps of the target task are randomly shuffled. That is, the target task might not be evolving smoothly with respect to the rotation degree.

- ImageCLEF-DA (Long et al., 2017): ImageCLEF-DA has three image classification tasks: Caltech-256 (C), ImageNet ILSVRC 2012 (I) and Pascal VOC 2012 (P). Following Wu and He (2022b), we generate the time evolving target task by adding random noise and rotation to the original images. For example, if we consider Caltech-256 (C) as the target task, we can generate a time-evolving target task by rotating the original images of Caltech-256 with a degree $O_d(j)$ ($j = 1, 2 \cdots, 5$ is the time stamp) and adding the random salt&pepper noise with the magnitude $O_n(j)$, i.e., $O_d(j) = 15 \cdot (j-1), O_n(j) = 0.01 \cdot (j-1), N = 4$.

Following Bobu et al. (2018) and Wu and He (2022b), we report both the classification accuracy on the newest target task (Acc) and the average classification accuracy on the historical target tasks (H-Acc) in the experiments. The comparison baselines we used in the experiments include: (1) static transfer learning approaches: SourceOnly, DAN (Long et al., 2015), DANN (Ganin et al., 2016), and MDD (Zhang et al., 2019); and (2) dynamic transfer learning: CUA (Bobu et al., 2018), GST (Kumar et al., 2020), L2E (Wu and He, 2022b), and our proposed L2S framework. For a fair comparison, all the methods use the same base models for feature extraction, e.g., LeNet for Rotating MNIST and ResNet-18 (He et al., 2016) for ImageCLEF-DA. In addition, we set $\eta = 1$, $\beta = 0.01$ and the number of inner epochs in $M_{ij}(\theta)$ as 1. All the experiments are performed on a Windows machine with four 3.80GHz Intel Cores, 64GB RAM and two NVIDIA Quadro RTX 5000 GPUs.

TABLE 2 Results of dynamic transfer learning on ImageCLEF-DA.

Methods	With continuous evolvement				With large distribution shift			
	I → C		I → P		I → C		I → P	
	Acc	H-Acc	Acc	H-Acc	Acc	H-Acc	Acc	H-Acc
SourceOnly	0.3125	0.4250	0.2812	0.3938	0.3125	0.4125	0.2187	0.2562
DAN (Long et al., 2015)	0.2500	0.4000	0.2187	0.2688	0.3750	0.3750	0.2500	0.2625
DANN (Ganin et al., 2016)	0.3125	0.4438	0.3125	0.4188	0.3125	0.4125	0.1875	0.2750
MDD (Zhang et al., 2019)	0.3437	0.4750	0.3125	0.4562	0.3125	0.4062	0.2500	0.3188
CUA (Bobu et al., 2018)	0.4063	0.5125	0.5312	0.5438	0.4375	0.4625	0.3437	0.4000
GST (Kumar et al., 2020)	0.5000	0.5312	0.4375	0.4312	0.2812	0.3062	0.2500	0.2562
L2E (Wu and He, 2022b)	**0.5625**	**0.6875**	0.5625	0.5875	0.3750	0.4812	0.3750	**0.4812**
L2S	**0.5625**	0.6125	**0.6562**	**0.6188**	**0.4375**	**0.5500**	**0.4375**	**0.4812**

The best results are indicated in bold.

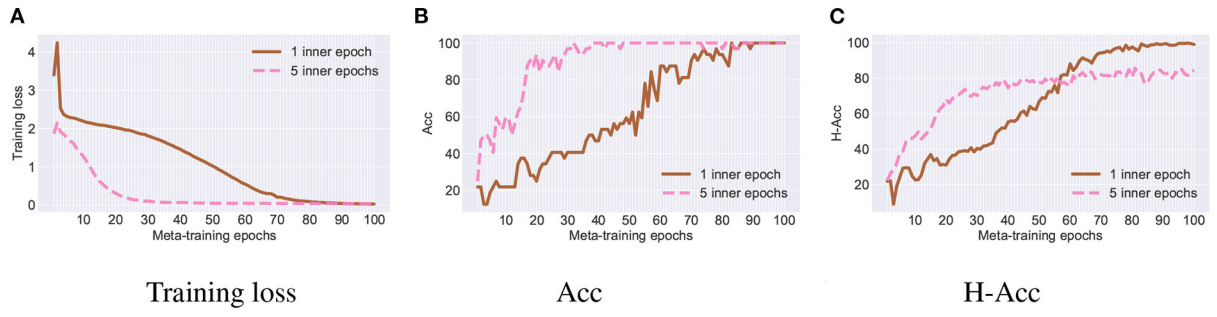

FIGURE 5
Ablation study with different number of inner epochs. **(A)** Training loss. **(B)** Acc. **(C)** H-Acc.

Input: A source task $\mathcal{D}^s$ (denoted as $\mathcal{D}_0^t$) and a dynamic target task $\{\mathcal{D}_j^t\}_{j=1}^N$, the newest target task $\mathcal{D}_{N+1}^t$.
Output: Prediction performance on the new target task $\mathcal{D}_{N+1}^t$.

Initialize the set of meta-pairs of tasks $\mathcal{S} = \emptyset$;
– – – – – – – – – – **Meta-training** – – – – – – – – – –
for $k = 1$ to N **do**
  Find all the candidate meta-pairs of tasks from $\mathcal{D}_0^t, \cdots, \mathcal{D}_k^t$;
  Estimate the sampling probability for these meta-pairs using Equation (4);
  Select a set of meta-pairs of tasks according to the sampling probability;
  Learn the model initialization $\tilde{\theta}^*$ via Equation (5);
  Generate the pseudo-label for $\mathcal{D}_k^t$;
end for
– – – – – – – – – – **Meta-testing** – – – – – – – – – –
Fine-tune on the newest target task $\mathcal{D}_{N+1}^t$ via Equation (6);
return Predicted labels on the newest target task $\mathcal{D}_{N+1}^t$.

Algorithm 1. Learning to Schedule (L2S).

6.2. Results

Figures 3, 4 show the distribution shift in the dynamic transfer learning tasks, where "S-T" denotes the distribution difference $d(\mathcal{D}^s, \mathcal{D}_j^t)$ between the source and the target at every time stamp and "T-T" denotes the distribution difference $d(\mathcal{D}_{j-1}^t, \mathcal{D}_j^t)$ of the target at consecutive time stamp. Here we use maximum mean discrepancy (MMD) (Gretton et al., 2012) to measure the distribution difference across tasks. We see that when the target task is continuously evolving over time, $d(\mathcal{D}_{j-1}^t, \mathcal{D}_j^t)$ is small. This enables gradual knowledge transferability in the target task. If there exists a large distribution shift at some times, i.e., $d(\mathcal{D}_{j-1}^t, \mathcal{D}_j^t)$ is large, the strategy of gradual knowledge transferability might fail. In Figures 3, 4, the large distribution shift happened in the time stamps 17–35 on Rotating MNIST and time stamp 1 on I → C/P.

Tables 1, 2 provides the experimental results of L2S as well as baselines on Rotating MNIST and Image-CLEF data sets. We have the following observations from the results. On the one hand, when the target task is continuously evolving over time, most dynamic transfer learning baselines can achieve satisfactory performance on both the newest and historical target tasks. The baseline GST (Kumar et al., 2020) fails on Rotating MNIST, because the self-training approach might be more likely to accumulate the classification error when the target task is evolving for a long time. On the other hand, the performance of CUA (Bobu et al., 2018) and L2E (Wu and He, 2022b) drops significantly when there is a large distribution shift within the target task at some time stamp. In contrast, by adaptively selecting the meta-pairs of tasks, the proposed L2S framework can mitigate the issue of the potential large distribution shift in the targe task. Specifically, compared to L2E (Wu and He, 2022b), L2S improves the performance by a large margin. This confirms the efficacy of the proposed progressive meta-pair scheduler.

6.3. Analysis

We provide the ablation study of our L2S framework with respect to the number of inner training epochs. The results on the newest target task of Rotating MNIST are shown in Figure 5, where we use 1 or 5 inner epochs for our meta-learning framework. We see that using more inner epochs can improve the convergence of L2S but it sacrifices the classification accuracy on the historical target task. This is because L2S with more inner epochs would enforce the fine-tuned model to be more task-specific. Thus, we set the number of inner epochs as 1 in our experiments.

7. Conclusion

In this paper, we study the problem of dynamic transfer learning from a labeled source task to an unlabeled dynamic target task. We start by deriving the generalization error bounds of dynamic transfer learning by assigning the meta-pairs of tasks with different weights. This allows us to provide the tighter error bound when there is a large distribution shift of the target task at some time stamp. Then we develop a novel meta-learning framework L2S with progressive meta-task scheduler for dynamic transfer learning. Extensive experiments on several image data sets demonstrate the effectiveness of the proposed L2S framework over state-of-the-art baselines.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JW and JH work together to develop a new theoretical understanding and algorithms for dynamic transfer learning. Both authors contributed to the article and approved the submitted version.

Funding

This work is supported by the National Science Foundation under Award Nos. IIS-1947203, IIS-2117902, and IIS-2137468 and Agriculture and Food Research Initiative (AFRI) Grant No. 2020-67021-32799/project accession no. 1024178 from the USDA National Institute of Food and Agriculture.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2022.1052972/full#supplementary-material

References

Acar, D. A. E., Zhu, R., and Saligrama, V. (2021). "Memory efficient online meta learning," in *International Conference on Machine Learning*, 32–42.

Acuna, D., Zhang, G., Law, M. T., and Fidler, S. (2021). "*f*-domain adversarial learning: theory and algorithms," in *International Conference on Machine Learning*, 66–75.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Mach. Learn.* 79, 151–175. doi: 10.1007/s10994-009-5152-4

Bitarafan, A., Baghshah, M. S., and Gheisari, M. (2016). Incremental evolving domain adaptation. *IEEE Trans. Knowl. Data Eng.* 28, 2128–2141. doi: 10.1109/TKDE.2016.2551241

Bobu, A., Tzeng, E., Hoffman, J., and Darrell, T. (2018). "Adapting to continuously shifting domains," in *International Conference on Learning Representations Workshop* (Vancouver, BC).

Chen, H.-Y., and Chao, W.-L. (2021). "Gradual domain adaptation without indexed intermediate domains," in *Advances in Neural Information Processing Systems, Vol. 34*, 8201–8214.

Finn, C., Abbeel, P., and Levine, S. (2017). "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning* (Sydney, NSW), 1126–1135.

Finn, C., Xu, K., and Levine, S. (2018). "Probabilistic model-agnostic meta-learning," in *Advances in Neural Information Processing Systems* (Montreal, QC), Vol. 31.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. (2019). "Online meta-learning," in *International Conference on Machine Learning* (Long Beach, CA), 1920–1930.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 2096–2030. doi: 10.1007/978-3-319-58347-1_10

Ghifary, M., Balduzzi, D., Kleijn, W. B., and Zhang, M. (2016). Scatter component analysis: a unified framework for domain adaptation and domain generalization. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1414–1430. doi: 10.1109/TPAMI.2016.2599532

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* 13, 723–773.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778.

Hoffman, J., Darrell, T., and Saenko, K. (2014). "Continuous manifold based adaptation for evolving visual domains," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 867–874.

Hospedales, T. M., Antoniou, A., Micaelli, P., and Storkey, A. J. (2021). Meta-learning in neural networks: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 14, 5149–5169. doi: 10.1109/TPAMI.2021.3079209

Kumar, A., Ma, T., and Liang, P. (2020). "Understanding self-training for gradual domain adaptation," in *International Conference on Machine Learning*, 5468–5479.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Liu, H., Long, M., Wang, J., and Wang, Y. (2020). "Learning to adapt to evolving domains," in *Advances in Neural Information Processing Systems, Vol. 33*, 22338–22348.

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning* (Lille), 97–105.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2017). "Deep transfer learning with joint adaptation networks," in *International Conference on Machine Learning* (Sydney, NSW), 2208–2217.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2018). Learning under concept drift: a review. *IEEE Trans. Knowl. Data Eng.* 31, 2346–2363. doi: 10.1109/TKDE.2018.2876857

Mancini, M., Bulo, S. R., Caputo, B., and Ricci, E. (2019). "Adagraph: unifying predictive and continuous domain adaptation through graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 6568–6577.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). "Domain adaptation: learning bounds and algorithms," in *22nd Conference on Learning Theory, COLT 2009* (Montreal, QC).

Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Shen, J., Qu, Y., Zhang, W., and Yu, Y. (2018). "Wasserstein distance guided representation learning for domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32* (New Orleans, LA).

Tang, S., Su, P., Chen, D., and Ouyang, W. (2021). Gradient regularized contrastive learning for continual domain adaptation. *Proc. AAAI Conf. Artif. Intell.* 35, 2665–2673. doi: 10.1609/aaai.v35i3.16370

Taufique, A. M. N., Jahan, C. S., and Savakis, A. (2022). "Unsupervised continual learning for gradually varying domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 3740–3750.

Tripuraneni, N., Jordan, M., and Jin, C. (2020). "On the theory of transfer learning: The importance of task diversity," in *Advances in Neural Information Processing Systems, Vol. 33*, 7852–7862.

Wang, H., Li, B., and Zhao, H. (2022). "Understanding gradual domain adaptation: improved analysis, optimal path and beyond," in *Proceedings of the 39th International Conference on Machine Learning* (Baltimore, MD), 22784–22801.

Wang, H., He, H., and Katabi, D. (2020). "Continuously indexed domain adaptation," in *Proceedings of the 37th International Conference on Machine Learning*, 9898–9907.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. (2020). "On the global optimality of model-agnostic meta-learning," in *International Conference on Machine Learning*, 9837–9846.

Wu, J., and He, J. (2020). Continuous transfer learning with label-informed distribution alignment. *arXiv preprint arXiv:2006.03230*. doi: 10.48550/arXiv.2006.03230

Wu, J., and He, J. (2022a). "Domain adaptation with dynamic open-set targets," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington, DC), 2039–2049.

Wu, J., and He, J. (2022b). "A unified meta-learning framework for dynamic transfer learning," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22* (Vienna), 3573–3579.

Wu, J., and He, J. (2021). "Indirect invisible poisoning attacks on domain adaptation," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1852–1862.

Wulfmeier, M., Bewley, A., and Posner, I. (2018). "Incremental adversarial domain adaptation for continually changing environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (Brisbane, QLD: IEEE), 4489–4495.

Yao, H., Wang, Y., Wei, Y., Zhao, P., Mahdavi, M., Lian, D., et al. (2021). "Meta-learning with an adaptive task scheduler," in *Advances in Neural Information Processing Systems, Vol. 34*, 7497–7509.

Zhang, Y., Liu, T., Long, M., and Jordan, M. (2019). "Bridging theory and algorithm for domain adaptation," in *International Conference on Machine Learning* (Long Beach, CA), 7404–7413.

Zhou, Y., Ma, F., Gao, J., and He, J. (2019b). "Optimizing the wisdom of the crowd: Inference, learning, and teaching," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, eds A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis (Anchorage, AK: ACM), 3231–3232. doi: 10.1145/3292500.3332277

Zhou, Y., Ying, L., and He, J. (2019a). Multi-task crowdsourcing via an optimization framework. *ACM Trans. Knowl. Discov. Data* 13, 1–26. doi: 10.1145/3310227

Zhou, Y., Yong, L., and He, J. (2017). "MultiCmbox 2: An optimization framework for learning from task and worker dual heterogeneity," in *Proceedings of the 2017 SIAM International Conference on Data Mining* (Houston, TX: SIAM), 579–587. doi: 10.1137/1.9781611974973.65