A Decade Survey of Transfer Learning (2010–2020)

Shuteng Niu, Yongxin Liu, Jian Wang, and Houbing Song, *Senior Member, IEEE*

***Abstract*—Transfer learning (TL) has been successfully applied to many real-world problems that traditional machine learning (ML) cannot handle, such as image processing, speech recognition, and natural language processing (NLP). Commonly, TL tends to address three main problems of traditional machine learning: (1) insufficient labeled data, (2) incompatible computation power, and (3) distribution mismatch. In general, TL can be organized into four categories: transductive learning, inductive learning, unsupervised learning, and negative learning. Furthermore, each category can be organized into four learning types: learning on instances, learning on features, learning on parameters, and learning on relations. This article presents a comprehensive survey on TL. In addition, this article presents the state of the art, current trends, applications, and open challenges.**

***Index Terms*—Transfer learning, machine learning, domain adaptation, distant domain transfer, cross-modality transfer learning.**

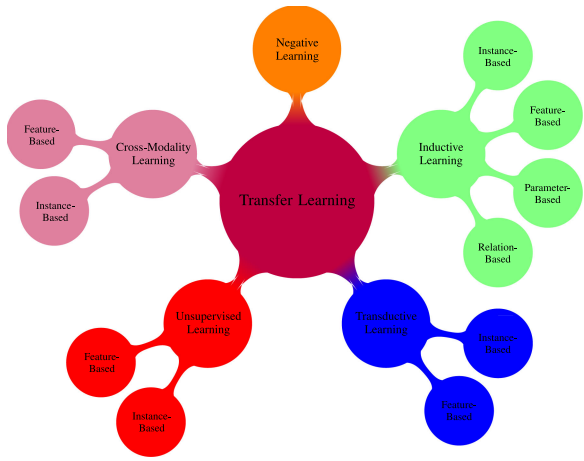


Fig. 1. Mindmap of Transfer Learning.

I. Introduction

RECENTLY, ML has made breakthroughs in a number of different fields, including but not limited to image processing, speech recognition, and natural language processing (NLP). With state-of-the-art performances, ML techniques have been applied to more and more real-world problems that traditional statistical learning methods cannot handle.

Commonly, traditional ML relies on a massive amount of training data. It assumes one critical condition: the training data and the testing data are drawn from the exact same distribution. However, this assumption does not always hold in many real-world problems. As such, most conventional ML algorithms usually suffer from three main difficulties: insufficient data, incompatible computation power, and distribution mismatch. First of all, various solutions have been proposed to address the first two problems, such as data argumentation, data synthesis, distributed learning, and cloud computing. However, each of these proposed solutions suffers from some adversities, such as regarding cost, efficiency, and security. Recently, transfer learning (TL) has been brought to our attention to deal with all three difficulties.

Manuscript received September 30, 2020; revised November 29, 2020 and January 18, 2021; accepted January 22, 2021. Date of publication January 26, 2021; date of current version March 5, 2021. This work was supported by the National Science Foundation under Grant 1956193. *(Corresponding author: Houbing Song.)*

The authors are with the Security and Optimization for Networked Globe Laboratory (SONG Lab), Department of Electrical Engineering and Computer Science, Embry-Riddle Aeronautical University, Daytona Beach, FL 32114 USA (e-mail: SHUTENGN@my.erau.edu; LIUY11@my.erau.edu; WANGJ14@my.erau.edu; h.song@ieee.org).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAI.2021.3053511.

Digital Object Identifier 10.1109/TAI.2021.3054609

Primarily, TL aims to solve the target task by leveraging the knowledge learned from source tasks in different domains, so it does not need to learn from scratch with a massive amount of data [1]–[3]. As such, TL first can address the most significant issue, insufficient well-labeled training data. Moreover, the time and computation resources required for training a model can also be greatly decreased since pre-learned knowledge from other domains and tasks can be reused. Furthermore, the distribution mismatch can cause significant performance degradation on ML models. TL can also address it by fusing knowledge from one or multiple different domains.

In this survey, the most representative works on TL in the past decade will be introduced and organized into different categories. Firstly, we categorize TL methods into two levels. As shown in Figure-1, in the first level, according to the availability of well-labeled data and the data modality in the source and target domains, it is categorized into five sub-fields: inductive TL, transductive TL, cross-modality TL, unsupervised TL, and negative TL respectively. Innovatively, each sub-field in the first level is again categorized into four different learning types: learning on instances, learning on features, learning on parameters, and learning on relations. Moreover, many successful real-world TL applications will also be introduced to emphasize TL's importance to the industry. And more, negative learning also plays a vital role in TL, which is an essential topic of TL but lacks attention. It is not studied by different learning types in the second level. Instead, it is discussed from two perspectives: problem definition and algorithms. In this survey, a number of state-of-the-art works on negative transfer will be discussed. Furthermore, open challenges and future research directions are also discussed in this survey.

TABLE I
COMPARISON OF RECENT SURVEYS ON TL

	Statistical	Deep Learning	Homogeneous	Heterogeneous	Negative	Cross-Modality	Applications
[2]	Yes	No	Yes	No	Yes	No	Yes
[4]	Yes	Yes	Yes	Yes	Yes	No	Yes
[5]	Yes	Yes	Yes	No	No	No	Yes
[6]	Yes	Yes	No	Yes	Yes	No	Yes
[7]	No	Yes	Yes	No	No	No	Yes
[8], [9], [10], [11], [12], [10]	Yes	Yes	No	No	No	No	Yes
Our Survey	Yes	Yes	Yes	Yes	Yes	Yes	Yes

TABLE II
TERMINOLOGY DEFINITION

	Domains	Tasks	Modalities
Inductive TL	Same	Same	Same
Transductive TL	Same	Different but related	Same
Unsupervised TL	Different but related	Different but related	Same
Cross-Modality TL	Different	Different	Different

TABLE III
TRANSFER LEARNING

Non-Cross-Modality	
Transductive Learning IV-A	
Feature-Based	[16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [3], [29], [30]
Instance-Based	[31], [32], [33], [34]
Inductive Learning IV-B	
Feature-Based	[35], [36], [37], [38], [39], [30], [27], [40], [41]
Instance-Based	[42], [43], [44], [45], [46], [47], [48]
Parameter-Based	[49], [50], [51], [52], [53]
Relation-Based	[31], [32], [33], [34]
Unsupervised Learning IV-D	
Feature-Based	[2], [54], [55], [56]
Negative Learning IV-E	
Problem Definition	[2], [57], [4], [58], [59], [60]
Algorithms	[61], [62], [63], [54]
Cross-Modality	
Cross-Modality Learning IV-C	
Supervised Target Data	[64], [65]
Semi-supervised Target Data	[66], [67], [68], [69]

Comparing with other recent surveys on TL, as shown in Table-I, we have several main improvements and contributions in this survey paper. The following outlines the main contributions of our survey:

- Introduce over 115 representative works from 2010 - 2020. Provide detailed explanations of each category's most famous works and discuss inter-connections of all works in each category.
- Discuss the most challenging topic, Cross-Modality TL, which has never been discussed in any previous surveys.
- Present deep insights to current challenges and frontier of TL applications.
- This survey can be used as a guideline for professionals to develop TL models.

Finally, the remainder of this paper is structured as follows: In Section-II, we introduce a number of recent surveys on TL, and demonstrate the improvements made by our survey. And then, in Section-III, we give an overview of the survey, and this overview can help professional to find well-suited methods for different situations quickly. Secondly, in Section-IV, we first review the most recent TL works. In-between, we also introduce some successful applications in industries. And then, we present the future trends and the open challenges in Section-V and Section-VI. Finally, we conclude the article in Section-VII.

II. RELATED WORK

In this section, as shown in Table-I, we review several surveys on TL in the past decade. Moreover, we demonstrate the main differences in our survey to distinguish it from other recently published works.

Recently, some surveys of TL with informative contents are provided for readers from both the academies and the industries. These surveys [2], [4]–[7], [9]–[15] categorize and review a wide range of TL techniques from different perspectives, such as algorithm types, applications, and the mixture of both.

First of all, we introduce some widely known surveys for TL algorithms. The survey [2] gives readers a brief overview and detailed explanations of representative TL algorithms from 2000 to 2010. However, this work does not cover several newly introduced TF disciplines, such as TL with artificial neural networks, heterogeneous TL, and TL with adversarial networks. In addition, the survey [4] gives attention to more recent TL topics that are not discussed in [2]. It introduces and summarizes a number of homogeneous and heterogeneous transfer learning algorithms from 2010 - 2015. More recently, another survey [5] gives special attention to homogeneous TL and reviews of state-of-the-art homogeneous TL algorithms and applications. It reviews homogeneous TL from two perspectives: the data and the model. However, some advanced topics are not covered in this survey, including but not limited to heterogeneous TL, reinforcement TL, and lifetime TL. Moreover, heterogeneous TL is specially discussed by the survey [6]. Recently, deep learning has received increasing attention from the TL community. A recent survey [7] focuses on TL with deep learning. It provides a formal definition of deep transfer learning and reviews current works in four deep TL disciplines: instance-based, mapping-based, network-based, and adversarial-based. Furthermore, there are some surveys [8]–[10], [10]–[12] particularly concentrate on TL applications in different fields: heath care systems, sentiment analysis, remote sensing, recommendation systems, and signal processing.

Our work covers the most recent topics in the past decade, such as TL with deep learning, TL with artificial neural networks, TL with statistical methods, TL with lifelong learning, and TL applications. Moreover, our survey also discusses the most challenging topic, cross-modality, and distant domain TL, which are not well-investigated in other surveys. Furthermore, detailed explanations of each type TF discipline's representative methods are provided for readers to have a better understanding. What is more, TL-related applications and current trends of TL are also discussed.

III. OVERVIEW

In this section, we give an overview of all methods that are discussed in the survey. As shown in Table-III, the table can be used as an index to help professionals to quickly find the works related to their specific interests. Moreover, it is also helpful for selecting appropriate methods to solve given TL problems.

There are three steps to find the most suited methods for a given TL problem. Firstly, it is essential to decide if the given

TABLE IV
TRANSDUCTIVE LEARNING

Transductive Learning	
Feature-Based	[16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [3], [29], [30]
Instance-Based	[31], [32], [33], [34]

TABLE V
INDUCTIVE LEARNING

Inductive Learning	
Feature-Based	[35], [36], [37], [38], [39], [30], [27], [40]
Instance-Based	[42], [43], [44], [45], [46], [47], [48]
Parameter-Based	[49], [50], [51], [52], [53]
Relation-Based	[31], [32], [33], [34]

TABLE VI
CROSS-MODALITY TRANSFER LEARNING

Cross-Modality Learning	
Supervised Target Data	[64], [65]
Semi-supervised Target Data	[66], [67], [68]

TABLE VII
UNSUPERVISED TRANSFER LEARNING

Unsupervised Learning	
Feature-Based	[2], [54], [55], [56]

TABLE VIII
NEGATIVE TRANSFER LEARNING

Negative Learning	
Problem Definition	[2], [57], [4], [58], [59], [60]
Algorithms	[61], [62], [63], [54]

TABLE IX
THE FRONTIER OF TL APPLICATIONS

Transfer Learning Applications		
Signal Processing	Transductive TL [88], [89], [90], [91], [92], [93], [94]	Distant Transfer [63], [61]
Sentiment Analysis	Inductive TL [95], [96], [97], [98]	Transductive TL [99], [100]
Health System	Inductive TL[101], [102], [103], [88], [14]	Transuctive TL[104]
CPS	Inductive TL[105], [106], [107]	Transuctive TL[108]

TABLE X
CHALLENGES IN APPLICATIONS

Challenges	Major Related Applications
Database for TL	Social Media, Online Shopping, Browsers, Web-based Applications
Perception TL	Virtual Assistant, Smart Homes, Smart Cities, Smart Wearings, Security Systems

problem is a regular TL task or a cross-modality. For example, from text to image is a cross-modality task, and from image to image is a conventional TL task.

For regular TL problems, there are four categories. The first three categories can be defined by the source domain's label availability and the target domain. Moreover, negative learning can be defined by measuring the statistical distance between the source domain and the target feature domain. For cross-modality TL problems, there are two categories defined by the label availability in the target domain.

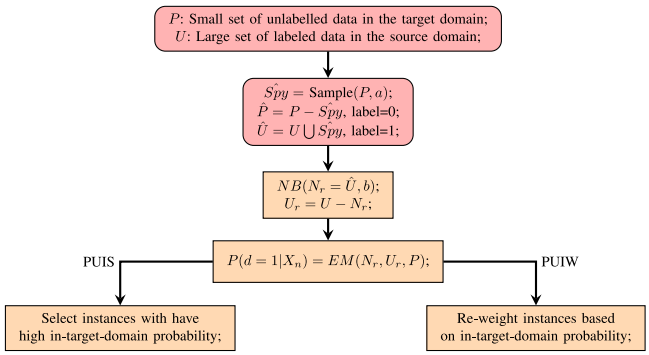


Fig. 2. PUIS & PUIW [31].

IV. STATE OF THE ART

This section presents the state-of-the-art of TL in the past decade.

A. Transductive TL

The definition of transductive learning [2] is: the tasks in the source and target domains are the same, but the domains may be different. Under this setting, the labeled data is only available in the source domain. Furthermore, there are two learning types in transductive transfer learning: learning on instances and learning on features. Moreover, the most widely known example of transductive learning is domain adaptation. In transductive TL, instance-based methods are not as popular as feature-based methods due to the limitations of its learning mechanism that is detailed in the following section. Therefore, the current mainstream of transductive TL is feature-based methods.

1) Learning on Instances: Primarily, algorithms of learning on instances are defined as transferring the knowledge in the source domain to the target domain by re-weighting or re-sampling source instances. Moreover, instance-based methods are built upon two strict assumptions: 1) the particular amount of training instances in the source domain are related to the target domain so that they can be reused, and 2) the conditional distributions of the source and the target domain are identical.

Importantly, not all the source data can be re-used for training the target model. Therefore, it is important to properly select samples that can benefit the task in the target domain. Firstly, [34] proposed a boosting method that leverages the concept of AdaBoost. Similarly, [31] proposed two novel approaches for instance re-weighting and instance selection based on the concept of PU learning and the in-target-domain probability. As shown in Figure-2, it first samples a small set $\hat{P}$ from unlabeled data P in the target domain as spies and labels all the instances $x \in P - \hat{P}$;n. as true. Then it labels $\hat{P} \bigcup U$ as false. A Naive Bayes (NB) classifier is then applied to $\hat{P}$ and U to identify a reliable negative set N_r based on the threshold b. The next step is to find the in-target-domain probability of $U_r = U - N_r$ by applying an Expectation Maximization (EM) algorithm. In Instance Selection (PUIS), the instances with higher in-target-domain probability are selected. Differently, Instance Weighting (PUIW) first calibrates the in target-domain-probability, and then use it as the sampling weights for training NB model.

However, methods similar to [31] are not efficient and heavily dependent on the pre-set values of the calibration parameters when the tasks have high-dimensional distributions. Moreover, some other instance-based adaptation models [32], [33] can handle tasks with have high-dimensional distributions. The core concept of this type of models is to adapt data in the source domain to the target domain by applying a logistic approximation.

More recently, [70] developed an instance-based multi-source transfer learning method based on the maximal correlation analysis [71]. Notably, it does not require the data from source domains to train a target domain model. Instead, it only requires the pre-trained source domain models to construct a set of distributed networks as a feature extractor for the target domain data. By doing this, the computation of the training is significantly reduced. What is more, a novel maximal correlation metric [72] was introduced to measure the distribution distance. More than that, as shown below, it also proposed four rules for designing algorithm-specific TL algorithms. The four rules are:

- Minimize the weighted empirical loss over source and target domains.
- Assign balanced weights to data points, as focusing too much on specific data points leads to over-fitting caused by perturbations in the training data.
- Assign more weight to the target sample, since target data will be used for testing.
- Assign weights such that the performance gap between the domains is small.

Moreover, it also proposed a novel algorithm called GapBoost, which adjusts the instance weight matrix by applying on a novel domain distance measurement, $Y - Discrepancy$:

$$dist_Y(D_S, D_T) = sup|L_{D_S}(h) - L_{D_T}(h)|, h \in H,$$

where h is the optimal chosen learning model during each iteration in the training stage.

2) Learning on Features: However, those required conditions of instanced-based algorithms do not always hold in many real-world problems [23], [24], [40]. Alternatively, feature-based methods have been developed to solve the issues. Firstly, [30] introduced the idea of transferable features for deep neural networks. In general, learning on features only needs a weaker hypothesis: the distributions of the target domain and the source domain are similar. Intuitively, it tends to minimize the distribution mismatch between the source domain and target domain by transferring or re-representing features to another space. Generally, there are two types of feature-based transductive learning methods: data-centered methods [16]–[19], [23] and subspace-centered methods [21], [22], [24]–[26].

Generally, data-centered methods are to discover a uniform transformation that can convert the data from the source domain and the target domain to a domain-invariant space so that the distribution mismatch can be minimized without losing original information. However, so it does not work well when the target domain and the source domain have a large discrepancy. Differently, subspace-centered methods try to reduce the domain shift by manipulating sub-spaces of the source domain and the target domain. To do this, we need to find the appropriate projections for the data in both domains.

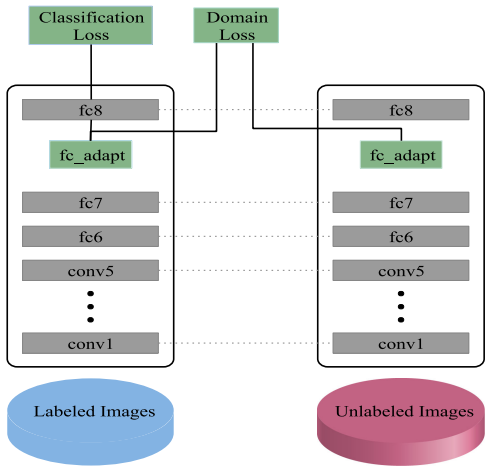


Fig. 3. Deep Domain Confusion [16], it is an AlexNet-based architecture with one adaptation layer and an additional domain confusion loss (MMD-based) was proposed to learn a semantically meaningful and domain invariant representation.

Firstly, the idea of adaptation layer was proposed by [17]. It introduced a modified feedforward neural network, Domain Adaptive Neural Network (DaNN), with one adaptation layer. Importantly, the loss function is contains two parts: the general loss and the MMD loss. Additionally, the MMD loss is used to evaluate the distribution mismatch between the source domain and the target domain. The model has produced better performance than similar models [1], [73]. However, it is a very shallow and simple model, so the performance is limited. Furthermore, several studies have approved the deep neural networks can learn much more transferable features, so we would like to benefit from the deeper features. To explore the potential of DaNN, a number of novel methods were proposed [16], [18]–[21]. As illustrated in Figure-3, Deep Domain Confusion (DDC) [16], an AlexNet-based [74] Convolutional Neural Network (CNN) with one adaptation layer and an additional domain confusion loss (MMD-based) was proposed to learn a semantically meaningful and domain invariant representation. Additionally, the evaluation metric can also be used to determine the position and the dimensionality of the adaptation layer. Furthermore, [22], [24] improved the performance of [16] by introducing weighted-MMD with weight regularizer. Moreover, [23] added another term, CORAL loss, to the regular loss function to produce even better results. In this method, CORAL loss, ℓ_{CORAL}, is defined as the distance between the second-order covariances of the source and the target features:

$$\ell_{CORAL} = \frac{1}{4\,d^2}\|C_S - C_T\|_F^2$$

However, the buried features in the deep layers could be highly task-specific, so that they cannot be safely transferred to new tasks. To solve this issue, another framework was proposed by [18]. It introduced a novel framework, deep adaptation networks (DAN), to enhance the feature transferability and reduce the domain shift. Differently, multi-kernel MMD is used to close the distribution mismatch between the source domain and the target domain, and multiple adaptation layers are applied to

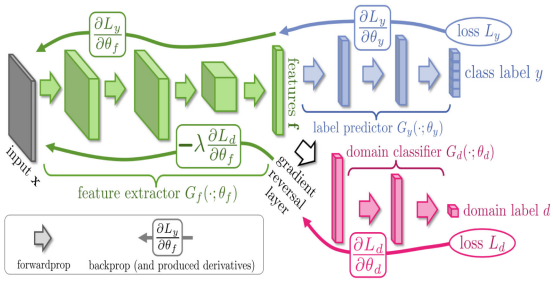


Fig. 4. Gradient Reversal [21], it has three components: a feature extractor (green), a label predictor (blue), and a domain classifier (blue).

improve the performance. As a classic example of multi-kernel MMD-based architectures, The Domain Adaptive Hash (DAH) network [26] combines hashing techniques and multi-kernel MMD. To the best of our knowledge, it is the first research that exploits the feature learning capabilities of neural networks to learn representative hash codes to address the domain adaptation problem. Particularly, hashing techniques can also convert the high dimensional data into binary codes, so it will be easier to access and store. In addition, there are more models [3], [27], [28] that have used adaptation layer. Especially, [29] is able to transfer across domains and tasks simultaneously.

Differently, [21] wishes to learn the underlying features that combine the discriminativeness and domain-invariance. The network architecture is shown in Figure-4. It has one feature extractor and two sub-classifiers. The underlying features can be learned by training two classifiers simultaneously, label predictor and domain classifier. The feature extractor can minimize the loss of the label predictor and maximize the loss of the domain classifier to make sure the features are domain-invariant. The loss function is constructed as:

$$E(\theta_f,\theta_y,\theta_d)=\sum_{i=1,d_i=0}^{N}L_y^i(\theta_f,\theta_y)-\lambda\sum_{i=1}^{N}L_d^i(\theta_f,\theta_d),$$

where L_y is the loss for label prediction, L_d is the loss for domain classification. However, the standard stochastic gradient descent does not fit this procedure because of the negative sign in front of the L_d loss. To solve this problem, gradient reversal layer (GRL):

$$R_\lambda(x)=x,$$
$$\frac{dR_\lambda}{dx}=-\lambda\mathbf{I},$$

was introduced to smoothly connect the feature extractor and domain classifier. Next, the GRL function is plugged into the loss function:

$$\widetilde{E}(\theta_f,\theta_y,\theta_d)=\sum_{i=1,d_i=0}^{N}L_y^i(\theta_f,\theta_y)+\sum_{i=1}^{N}L_d^i(G_d(R_\lambda(G_f(x_i;\theta_f));\theta_d),y_i) \quad (1)$$

B. Inductive TL

Unlike transductive learning, inductive learning is defined as: the tasks in the source domain and the target domain are different regardless if the domains are the same or not. Under this setting, the well-labeled data is usually available in the target domain, no matter the well-labeled data is available or unavailable in the source domain. Particularly, the main focus is on the former. In this case, inductive learning is similar to multi-task learning, but it only concentrates on the target task. Differently, when there is no labeled data in the source domain, inductive learning is close to self-taught learning. The information is hidden in the source domain, so it cannot be used directly. Commonly, inductive TL aims to develop a target model with a small set of well-labeled data in the target domain.

Additionally, there are four learning types in inductive learning: learning on instances, learning on features, learning on parameters, and learning on relations. Furthermore, the first three types of methods are the mainstream in inductive learning, while relation-based methods are not very common.

1) Learning on Instances: Generally, the training data in the source domain are more or less out-dated, and processing new data is very costly. Inductive TL aims to train an accurate model with only a tiny amount of well-labeled training data in the target domain. Moreover, the key of this type of methods is finding which part of the old data can be adapted to train a new model in the target domain. One of the most famous instance-based methods in inductive learning is TrAdaBoost [42], an AdaBoost [43]-based transfer learning algorithm. Conceptually, it extracts useful information in the source domain by iteratively re-weighting the source domain instances. Firstly, it employs a few labeled new data, called same-distribution data T_s, to evaluate the value of each old the old data in the source domain. Furthermore, the instances with low value are classified as diff-distribution data T_d. And then, it combines T_d, T_s, and unlabeled data S to train a new model for the target task. However, the re-weight procedure of TrAdaBoost is not the same as AdaBoost. Additionally, it increases the weights of incorrectly predicted instances in T_d, while decreases the weights of correctly predicted instances in T_s. Similarly, [72] proposed GapBoost, a novel multi-source boost method for transfer learning.

Recently, several algorithms inspired by TrAdaBoost have pushed the performance to a new level. Firstly, one of the shortcomings of TrAdaBoost is only using one type of base learner to train the model in the target domain, but there might be other base learners that can give better performance. To address this issue, [44], [45] choose to employ different base learners to improve the performance on specific tasks. Secondly, the original TrAdaBoost algorithm only uses one source domain for the knowledge transfer. However, the knowledge is not always enough from a single source domain. In order to overcome this shortcoming, [46], [75], [76] take advantage of combining multiple source data sets to avoid negative learning. Additionally, [76] can decide which sources are helpful to build the model in the target domain by iteratively performing two types of boosting: 1) individual boosting for instances and 2) task-based boosting. It increases the weights of incorrectly predicted instances, and

Algorithm 1: TrAdaBoost.

Input: Two labeled training sets T_d and T_s;
The unlabeled testing set S;
Initialize: Learner, F;
The number of iterations, N;
Weight Vector, W^1;
for $t = 1, ..., N$ **do**

1. Set $P^t = w^t / \sum_{i=1}^{N} w_i^t$.
2. Apply **Learner**, $F_t(X) = Y$.
3. Calculate the error:

$$\epsilon_t = \sum_{i=n+1}^{n+m} \frac{w_i^t |h_t(x_i) - c(x_i)|}{\sum_{n+1}^{n+m} w_i^t}.$$

4. Set $\beta_t = \epsilon_t/(1-\epsilon_t)$ and $\beta = 1/(1+\sqrt{2\ln n/N})$.
5. Update the new weight vector:

$$w_i^t + 1 = \begin{cases} w_i^t \beta^{|} h_t(x_i) - c(x_i)| \\ w_i^t \beta^{-} |h_t(x_i) - c(x_i)|. \end{cases}$$

end
Output: $F_t(x) =$

$$\begin{cases} 1, & \prod_{t=[\frac{N}{2}]}^{N} \beta_t^{-h_t(x)} \geq \prod_{t=[\frac{N}{2}]}^{N} \beta_t^{-\frac{1}{2}} \\ 0, & otherwise \end{cases}$$

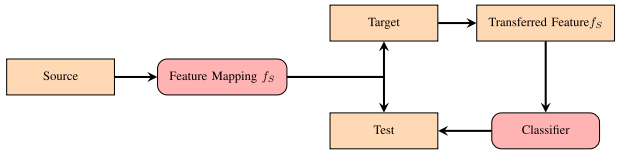


Fig. 5. Feature-Based Inductive Learning [79], it requires a large set of labeled data from one or multiple source domains, and a small amount labeled data from the target domain. The core idea is to train three separate model by augmenting the original data into three sets, namely, source-specific, target-specific, and general-specific.

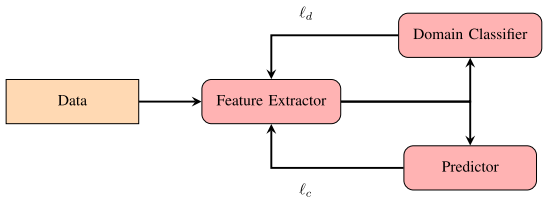


Fig. 6. GANs-Based TL.

it also performs a task-based boosting that can enhance the instances from the tasks that have higher transferability. Unlike TrAdaBoost, it keeps all the base learners can improve the performance of the model because the early iterations fit the majority of the data while the later iterations focus on more in-depth details. Furthermore, there are also researches [47], [48] that improve the model with dynamic weight update methods.

Overall, re-weighting instances iteratively is a proven way to enhance inductive learning models' performance, yet some other researchers hold different opinions. Commonly, certain parts of the differently distributed data T_d could help training the model in the target domain, yet certain parts could also be harmful. Moreover, there are no simple methods to measure the transferability of the source data sets accurately. Therefore, some algorithms [77], [78] intend to remove all the different distribution data instead of assigning small weights to them.

2) Learning on Features: Commonly, feature-based inductive transfer learning algorithms [27], [30], [35]–[40] wish to extract shared features to minimize domain divergence and model error. According to the types of source data sets, feature-based algorithms can be classified into two categories: supervised and unsupervised. Firstly, supervised algorithms [27], [30], [35]–[39] are similar to multi-task learning, which combines a sufficient amount of labeled source data and a tiny amount of labeled target data to train a high-quality model in the target domain. However, multi-task learning tends to learn all the tasks simultaneously, while inductive transfer learning only focuses on the target task. Differently, unsupervised algorithms [40], [72] are more powerful but difficult to train.

Primarily, most feature-based inductive transfer learning methods focus on finding domain-invariant features. In other words, the problems can be converted into how to effectively extract features that can reduce the divergence between the source domain and the target domain [2]. [35] introduces a simple, fully supervised approach with feature-augmentation. Firstly, it requires a large set of labeled data from one or multiple source domains, and a small amount labeled data from the target domain. And then, it trains three separate model by augmenting the original data into three sets, namely, source-specific, target-specific, and general-specific. Additionally, three sets of weights of those data sets are denoted as W_s, W_t, W_g. Moreover, W_s represents the sum of the "source" and "general" features, W_t represents the sum of the "target" and "general" features. And the feature-augmented weights are regularized by $|W_g|^2 + |W_s - W_g|^2 + |W_t - W_g|^2$. Finally, minimizing the sum of the equation will find the features that can minimize the divergence. Moreover, as shown in Figure-5, [79] proposed a framework to justify the effectiveness of feature-based inductive transfer learning. Firstly, it constructs a feature mapping, F, for the source domain data. Then use this mapping to transfer the target domain data to the same feature space. After that, it trains a discriminative classification model based on the feature extracted by F. Besides, the mapping learned in the first step will also be used to convert the test data into the same feature space as the training data.

Recently, several works [27], [30], [39], [40] have evaluated the combination of GANs and transfer learning. Initially, this kind of methods aim to free human from hand-designing networks for extracting shared features. For example, [39] aims to find features that are 1) discriminative for the main learning task in the source domain and 2) domain-invariant by implementing the idea of GANs. Moreover, these features are considered ideal for cross-domain transfer when we cannot identify the original domain of the inputs [30]. As shown in Figure-6, the model includes three main components, domain classifier G_f, predictor P, and feature extractor G_f. The final goal is to learn the mapping (M) to predict unknown instances in the target domain with low risk. Furthermore, the risk is defined as follow:

$$R_{D_T}(M) = Pr_{(x,y)\, D_T(M(x) \neq y)}$$

where D_T represents the target domain, and M represents the mapping from the features to the labels.

Similar to the typical GANs model, domain classifier and predictor will be adversarial to each other. As shown in Figure 6, the parameters of the domain classifier are trained to minimize the loss during the training. The feature extractor parameters are

optimized to minimize the loss of the predictor ℓ_c and maximize the loss of the domain classifier ℓ_d. Therefore, the loss of the model is constructed by two terms:

$$\ell = \ell_c(D_s, y_s) + \lambda \ell_d(D_s, D_t)$$

where D_s represents the source domain, D_t represents the target domain, and λ is the learning coefficient.

3) Learning on Parameters: Generally, parameter-based approaches [49]–[53] are based on the assumption that there are shared-parameters in models from source domains and the target domain. Thus, this type of methods are not suitable for the cases with a significant domain shift. Under this setting, parameter-based methods can be easily derived from multi-task learning methods. However, multi-task learning is usually focused on learning all the tasks simultaneously, while parameter-based transfer learning is only focused on optimizing the target task. Thus, the loss functions for all the tasks are the same in multi-task learning, but the loss function in the target domain has greater weights in the transfer learning.

Firstly, [49] introduced a decision tree embedded transfer learning framework. TransEMDT (Transfer learning EMbedded Decision Tree) aims to address supervised transfer learning problems. As shown in Figure-, it first trains a decision tree with the source data (DT_S). Secondly, it feeds a small amount of the labeled target data into DT_S, and the prediction is used as initial clusters for K-Means model. After that, the parameters of DT_S is updated. Then the previous steps will be repeated until it converges. Finally, the output will be the decision tree for DT_T. Similar to TransEMDT, [51] proposed another framework, TransRKELM (Transfer learning Reduced Kernel Extreme Learning Machine), which uses RKELM to build an initial activity recognition model. Furthermore, several algorithms [52] have achieved promising performance by modifying SVM (Support Vector Machine). Typically, they assume that weight vectors of SVM contains two components: $W = W_S + W_D$, where W_S represents weight vectors that are shared across the source and the target domains, while W_D represents domain-specific weight vectors. In general, the traditional discriminative query strategy results in poor performance when there is a significant distribution mismatch between the source domain and the target domain. Some studies [52], [53] applied the generative query strategy to overcome this shortcoming. Moreover, [53] extended binary learning method to multiclass problems by implementing the one-vs-all approach. Furthermore, [50] presented Multilinear Relationship Networks(MRN). It can prevent negative transfer in the feature layers by jointly learning transferable parameters and multilinear relationships.

4) Learning on Relations: Comparing to other topics in inductive TL, relation-based transfer learning is not very popular. Unlike the other three types of learning methods, relation-based transfer learning methods do not assume the source data and the target data to be independent and identically distributed (i.i.d). This makes relation-based methods much more flexible and robust than traditional methods. However, there are not many studies on this topic in recent years. Moreover, most of this type of algorithms are built based on statistical learning techniques. The idea behind relation-based transfer learning is that similar relations exist in different domains. For example, the data in the source domain contains images of a professor giving a lecture to students, and the data in the target domain contains images of a manager giving a speech to employees. Although two sets of images describe different objects, they have the same relation.

Some studies [80]–[82] have proposed to use Markov Logic Networks. As shown in Figure-, the Markov Logic Network can be demonstrated by finding similar relationships from two different domains to construct a mapping from the source domain to the target domain.

C. Cross-Modality Transfer Learning

Commonly, most TL algorithms require more or less the connection in feature spaces or label spaces between the source and the target domain. In other words, knowledge transfer can only be performed when the source data and the target data are in the same modality, such as image, audio, and text. Unlike all other TL methods, Cross-Modality Transfer Learning (CMTL) is one of TL's most challenging topics. It assumes that the source and the target domain's feature spaces are entirely different, such as from text to image, from audio to text, and from image to audio. Moreover, the label spaces between the source and the target domain can also be different.

Intuitively, CMTL is inspired by humans' ability to generalize knowledge from one subject to another by building a bridge with other subjects. For example, a child who has read an article with descriptions of monkeys, and he has never seen any monkeys or images of monkeys. However, it is very possible that the child can recognize a monkey based on that article's knowledge. In this case, a child can transfer the knowledge from text data to image data using knowledge in other different domains. Theoretically, two seemingly unrelated domains can be connected by one or multiple bridge domains with overlapping semantic information. However, this type of learning behavior is difficult for machines to mimic due to the challenge in selecting appropriate intermediate domains as the bridge. Moreover, there are two types of CMTL algorithms: CMTL with Supervised Target Data and CMTL with Supervised Target Data.

1) CMTL With Supervised Target Data: This section discusses several text-to-image (TTI) DDTL methods, which require a small set of labeled image target data. Importantly, image classification tasks currently have two challenges: 1) labeled image data is relatively scarce and expensive to collect, and 2) features of image data lack semantic meaning for class prediction as they represent visual features rather than conceptual ones. Moreover, labeled text data is often more accessible than labeled image data, and text features have more semantic meaning for predicting a class label.

Firstly, Translated Learning via Risk Minimization (TLRisk) was introduced by [64]. It proposed an asymmetric architecture to map the features in the source domain to the target domain. Moreover, it uses a language model [83] and the nearest neighbor method to connect the text source data and the image target data. Moreover, for a smooth feature transition, it builds a translator by applying the Markov chain. The source features and the target features are modeled by two different Markov chains, which can be bridged with intermediate data. In other words, the translation is done by learning a probabilistic model that

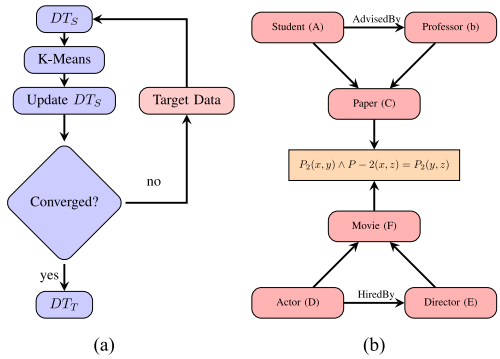


Fig. 7. (a) TransEMDT [49], it first trains a decision tree model based on the source data (DT_S). Secondly, it feeds a small amount of the labeled target data into DT_S, and the prediction is used as initial clusters for K-Means model. (b) Markov Logic Network [80]–[82], the Markov Logic Network can be demonstrated by finding similar relationships from two different domains to construct a mapping from the source domain to the target domain.

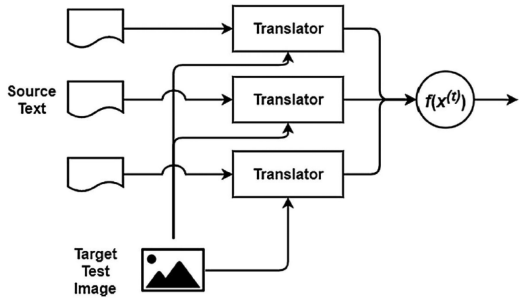


Fig. 8. Text-to-Image [65], CMTL transfers between knowledge between text files and images with multiple translators.

uses cooccurrence data as a bridge between the source and target feature spaces. Finally, it uses a variant of the risk minimization model to produce the final label prediction. This method outputs promising results that are better than the baseline model trained on only target data. However, the computational cost of TLRisk is very expensive due to the risk function estimation and dynamic programming.

To decrease the computational cost, [65] proposed another method for text-to-image (TTI) classification. In this study, the source domain is text data, and the target domain is image data. This method implements a novel transition method, translator, to build a bridge from text to images. It requires labeled source text data, text-image cooccurrence data, and a small amount of labeled image target data. This method uses TL to exploit such text data to improve image classification. Therefore, this problem is converted to how to relate the text to semantic knowledge transfer images. Moreover, this method uses a text-image cooccurrence matrix that contains images and the text that occurs with them on the same webpage. Cooccurrence information is effective because of the assumption that the text around an image describes the concepts in such an image. This cooccurrence information is relatively inexpensive to collect and serves as a bridge to learn the correspondence for translating the semantic information between the source text and the target image. This translation is achieved by the form of a feature transformation called a "semantic translator function." This translator takes the source, the target, and the cooccurrence data and learns the correspondence between the source text and the target images through the cooccurrence bridge. Each translator for the source text contains a "topic space," a common subspace associated with the translation data. As shown in Figure-8, there are a number of translators combined to form the final decision function $f(x^{(t)})$. Furthermore, this method bypasses the performances conducted by [64], [66] and other benchmark models trained with only target data, and it yields state-of-art accuracy with only a little target training data.

2) CMTL With Semi-Supervised Target Data: Unlike CMTL with supervised target data, several methods can take labeled and unlabeled target data to improve the classification performance.

Firstly, [66]proposed a heterogeneous TL for Image Classification (HTLIC) method that can take in semi-supervised source data and target data. Moreover, it aims to enhance a target image classification task with limited labeled data by exploiting semantic knowledge derived from unlabeled text documents and unlabeled annotated images from an auxiliary source. The unlabeled auxiliary data is relatively inexpensive to collect and it can enhance target image classification performance. It aims to find the relationship between unlabeled source text data and the semi-supervised image target data using auxiliary data with related semantic information. Furthermore, the connection is discovered using a two-layer bipartite graph where the top layer represents the relationship between the images and the tags, while the bottom layer represents the relationship between the tags and the documents. The feature space gap between the source domain and the target domain can be reduced. Moreover, more shared semantic information can be discovered with this bridge in low-level features with semantic analysis [84]. Unlike previous methods, HTLIC does not use a Markov chain to achieve the classification task. Instead, it applies traditional support vector machines (SVMs) [85] to make the final predictions. As the main improvement of this method, it proposed an efficient way to utilize semi-supervised target data to produce promising classification accuracy.

Furthermore, [79] first introduced the idea of using co-occurrence information between two different domains. And then, [86] proposed Co-occurrence Transfer Learning (CT-Learn) for knowledge transfer between text data and image data. More importantly, it enables the knowledge transfer from multiple domains, significantly improving the target classification accuracy with appropriate source domain selection. Unlike the previous methods [66], CT-Learn first uses the co-occurrence information between the text data and image data to create a joint transition probability matrix P:

$$P = \begin{bmatrix} \lambda_{1,1}P^{(1,1)} & \lambda_{1,2}P^{(1,2)} & \dots & \lambda_{1,N}P^{(1,N)} \\ \lambda_{2,1}P^{(1,1)} & \lambda_{2,2}P^{(2,2)} & \dots & \lambda_{2,N}P^{(2,N)} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \lambda_{N,1}P^{(N,1)} & \lambda_{N,2}P^{(N,2)} & \dots & \lambda_{N,N}P^{(N,N)} \end{bmatrix}. \quad (2)$$

This matrix is constructed using intra-relationships and inter-relationships for all the co-occurrence, labeled, and unlabeled instances across both domains. Moreover, the intra-relationships

are calculated by the affinity of the intrinsic manifold structure between the ith domain, and the inter-relationships are calculated by using the co-occurrence information. The diagonal elements represent intra-relationships, and other elements indicate inter-relationships between the ith and the jth domains. The weights λ decide the amount of transferable knowledge between domains, which shares a similar idea of learning rate in artificial neural networks. Furthermore, after extract the inter-relationships and the intra-relationships, it creates a coupled Markov chain based on a random walk with a restart. Different from TLRisk [64], CT-Learn applies a variant of regular Morkov chain to adapt multiple source domains. Moreover, most previous methods can only handle binary classification problems, but CT-Learn can deal with binary and multi-class classification problems. Finally, CT-Learn performed the highest accuracy on most benchmark data sets.

D. Unsupervised Transfer Learning

Primarily, the idea of transfer learning was proposed to solve the issue of lacking data. Moreover, many transfer learning methods have successfully generalized machine learning techniques to practical and performance-critical problems. However, most algorithms are focused on supervised cases and semi-unsupervised cases. In general, supervised algorithms cannot deal with cases where we do not even have enough labeled in the source domains.

Conceptually, unsupervised TL is defined as no labeled data in both the source domain and the target domain. This type of methods are beneficial to tasks that are unique and special, so a sufficient amount of labeled data from both the source domain and the target domain are not accessible. However, researchers have not favored this topic due to some barriers that make it difficult to apply to real-world tasks. Generally, there is only one sub-field under this setting: feature-based learning. Additionally, unsupervised transfer learning is also termed as self-taught learning by many researchers and scholars.

1) Learning on Features: Firstly, a few methods [54], [55] for clustering and dimensional reduction problems were summarized by [2]. The concept of Self-taught Clustering (STC) was introduced by [55], which aims to perform clustering on a small set of unlabeled target data with the help of a sufficient amount unlabeled in the source domains. In theory, STC tends to convert data sets in different domains into a common feature space, which can utilize the source data to cluster the target data. Moreover, proposed Transferred Discriminative Analysis (TDA) was proposed by [54]. It can generate pseudo-class labels for the target data by applying clustering methods.

Furthermore, a novel self-taught learning algorithm was introduced by [56]. It uses sparse coding to construct higher-level features using the unlabeled data. Moreover, this algorithm had been shown to improve the performance of classification tasks significantly.

E. Negative Transfer Learning

Certainly, transfer learning has successfully solved the issue of lacking training data in many real-world applications. However, it also has one shortcoming: negative transfer. Commonly, negative transfer occurs when transferring too much unrelated knowledge from the source domains. Despite its pervasiveness, negative transfer is usually described in an informal manner, lacking rigorous definition, careful analysis, or systematic treatment. Firstly, there are numerous survey papers [2], [4], [57] have discussed this issue in many TL disciplines. Furthermore, some researches [58]–[60] have recognized it in many real-world applications. In this section, we introduce some of the works that address negative learning.

First of all, typical TL assumes that the target domain and the source domain are different but related, so some common instances or features can be transferred between different domains. However, it limits TL from being applied to cases where the source and the target are very loosely connected. To address this issue, some works focus on transferring knowledge between two distant domains. Firstly, an instance-based algorithm [61], transitive transfer learning (TTL). It transfers knowledge between text data in the source domain and the image data in the target domain by using annotate image data as the knowledge bridge. However, this algorithm is very situational and case-dependent. Moreover, another feature-based method [62] was proposed to deal with scarce satellite image data. It predicts the poverty based on the daytime satellite image by transferring knowledge of an object classification task with the help of some nighttime light intensity information as a bridge. The main contribution of this method is to use similar data with different conditions to connect two different domains. Moreover, an instance-based distant domain transfer learning (DDTL) algorithm [63] uses several intermediate domains to bridge the source and the target. More specifically, it first uses an auto-encoder pair to select instances from the source domain and the intermediate domains, and it also learns high-level representations for data in different domains. After that, it trains a CNN model by using the selected instances and representations. Importantly, this method can be simply generalized to different tasks and produce fairly decent results. However, there are some challenges need to be addressed. Firstly, most chosen instances are from the intermediate domains and only a little from the source domain. Furthermore, it makes the source data seem unnecessary. The second, it assumes that there is a sufficient amount of intermediate domain data so we can find enough samples to build the bridge connecting the source and the target domains. In some cases, enough intermediate domains might not be accessible.

Furthermore, a study [87] first derived a novel definition of negative from three different perspectives, the chosen model, the divergence between the joint distributions, and the size of labeled target data, respectively. More importantly, it proposed a new term, negative transfer gap (NTG), to quantify the effect of negative transfer. It then introduced a novel GANs-based instance re-weighting algorithm to select useful samples from the source domain.

V. Frontier of TL

In this section, we present the current trends in TL from two aspects: TL algorithms and TL applications. For TL algorithms, we introduce several fields in TL that attract most attention. For TL applications, we demonstrate various applications spanning

multiple TL disciplines. Moreover, the main attention of the algorithm level is in solving the issues of insufficient data and distant domain transfer by conducting experiments that usually step ahead of making real productions. Therefore, assumptions made in experiments do not always hold in real-world problems. Differently, real-world applications focus more on applying TL models with stable and promising performances, so methods with pre-assumptions cannot be used.

A. Frontier of Transductive TL

First of all, domain adaptation, a sub-field of transductive TL, is the most active area. It tends to solve problems where only have a sufficient amount of labeled source data and unlabeled target data for the training process. Therefore, domain adaptation methods can be categorized into a cluster of semi-supervised learning algorithms. Moreover, this semi-supervised manner gains more focuses than other TL topics do. Currently, existing domain adaptation algorithms aim to close the marginal distribution distance or conditional distribution distance in two ways: symmetrical training and asymmetrical training. The first, symmetrical training [16], [23] means that there are two models with identical structures for the source and the target domains. It is commonly applied to feature-based algorithms. The advantages of symmetrical training are: 1) easy to train, 2) fast convergence, and 3) robustness with small source data sets. However, it also suffers from a significant shortcoming: performance decrease due to large domain discrepancy. Moreover, asymmetrical training [61], [63] is related to the cases where the structures of the source model and Target model are not identical but have some common layers. In general, it is applied in instance-based domain adaptation algorithms. Furthermore, it can handle a large domain shift by selecting statistically similar instances from multiple source domains. However, with multiple source domains, asymmetrical training suffers from difficulties in the training and non-convergence.

Moreover, there are two common learning types of domain adaptation algorithms: feature-based and instance-based. Feature-based is the mainstream in the domain adaptation area since there is no labeled target data. Generally, feature-based methods aim to utilize all training samples from the source and the target by extracting shared features or closing the feature distribution distance. To extract common features, most algorithms first calculate the distance between low-level features from the source domain and the target domain with a distribution distance metric through each iteration. The next step is to select or re-weight the features based on the distribution distance to learn high-level feature combinations. Similarly, some feature-based algorithms tend to discover more shared features by converting features from different domains into a novel feature space where the distance of different features is small. Feature-based methods can carry out state-of-art performance when the source and the target have strong connections. However, the performance can drop if there is only a small amount of similar data samples across domains because a large number of different samples can overfit the model.

Differently, instance-based algorithms aim to select similar instances from different domains to ensure a safe and quality knowledge transfer. Combining instance re-weighting and distribution distance metric is the most commonly used technique in instance-based methods. Also, there are two different types of instance re-weighting: soft re-weighting and hard re-weighting. Firstly, soft re-weighting does not eliminate any instance. Instead, it just sets the weights of dissimilar instances to extremely small values. On the contrary, hard re-weighting eliminates all dissimilar samples by setting the weights to zero. With the selection procedure, training samples have a more reliable connection, and they can avoid the performance drop due to large domain discrepancy. Besides, instance-based methods can output relatively more stable performance. However, the performance can be disappointing when the volume of the source domain data is small because the number of selected instances can be insufficient if the domain distance is far.

B. Frontier of Inductive TL

Generally, there are two main types of inductive TL learning algorithms: multi-source TL and self-taught TL. In common, both learning algorithms require labeled target data for the training process. Moreover, multi-source learning also needs labeled source data, while self-taught does not rely on labeled source data. Furthermore, multi-source learning attracts more attention due to its stable performance.

The main idea of multi-source learning is to take the advantage of multiple source domains. It is difficult to extract enough shared information from a single source domain in real-world problems due to the distribution discrepancy. Therefore, we aim to utilize multiple source domains to discover common features from each source domain and combine them to develop a source domain model. Moreover, this type of algorithms are usually stable and robust, but they are also computationally expensive due to the quantity of data from various domains. Under the setting of this type of algorithms, instance-based learning methods are more preferred than feature-based algorithms because the number of source training samples is sufficient for the training process. Furthermore, multi-source TL is closely related to supervised multi-task learning, another favored non-transfer ML technique. They both utilize multiple data sets from different domains and tasks. However, multi-task learning aims to improve the models in all different domains by sharing data sets. Differently, multi-source TL only focuses on the model in the target domain for the target task. Therefore, multi-task learning achieves better overall performances for multiple domains, and multi-source learning carries out better performance for a model in a specific domain.

Unlike multi-source TL, self-taught TL only requires labeled data from the target domain, which is more powerful but more costly and challenging to train. Moreover, feature-based learning methods and instance-based methods are both available in self-taught TL. In feature-based methods, unsupervised feature construction is required since there are no labels for the source domain data. The most commonly used unsupervised feature construction is sparse coding, which can be treated as a two-step minimization problem. In instance-based methods, the original TrAdaBoost [75] is the cornerstone of many advance self-taught TL algorithms, including but not limited to multi-source

TrAdaBoost, weighted TrAdaBoost, and multi-class Boost. Furthermore, instance-based methods are generally easier to train because the convergence of unsupervised feature construction is not always guaranteed.

C. Frontier of Distant Domain TL

Recently, insufficient training data and domain distribution mismatch have become the two most difficult challenges in ML. To address these two issues, TL has attracted more attention due to its training efficiency and domain shift robustness. However, transfer learning also suffers from a critical issue, negative transfer [59]. It significantly limits the use and performance of transfer learning. This section introduces some related works in three fields: conventional transfer learning, DDTL, and multi-task learning.

Firstly, TL aims to find and transfer the common knowledge in the source domain and the target domain. Furthermore, a research [27] expands the use of TL from traditional machine learning models to deep neural networks. Typically, there are two types of accessible TL: feature-based and instance-based. Moreover, both types focus on closing the distribution distance between the source domain and the target domain. In instance-based algorithms, the goal is to discover source instances similar to target instances to eliminate the highly unrelated source samples. Differently, feature-based algorithms aim to map source features and target features into a common feature space where the distribution mismatch is minimized. However, both of them naturally assume that the source domain and the target domain share a reasonably strong connection. Unlike conventional transfer learning, our work can transfer knowledge between different domains and tasks that are not closely related.

Secondly, most DDTL algorithms are similar to multi-task learning [109], which also benefits from shared knowledge in multiple different but related domains. Generally, multi-task learning tends to improve the performance on all the tasks. Differently, DDTL only focuses on using the knowledge in other domains to improve the target task's performance on the target domain.

Lastly, most previous studies of DDTL focus on instance-based methods and tend to take advantage of massive related source data. Firstly, [61] introduced an instance-based algorithm, transitive transfer learning (TTL). It transfers knowledge between text data in the source domain and the image data in the target domain using annotate image data as a bridge. However, this algorithm is highly case-dependent and unstable on performance. Similarly, another instance-based algorithm was introduced by [63]. It proposed a novel instance selection method, Selective Learning Algorithm (SLA). Moreover, SLA can select helpful instances from many unrelated intermediate domains to expand the volume of the source data. However, this algorithm mainly aims to handle binary classification problems. Furthermore, a feature-based method [62] can deal with scarce satellite image data. It predicts the poverty based on the daytime satellite image by transferring knowledge learned from an object classification tasks with the help of some nighttime light intensity information as a bridge. However, this method heavily relies on a massive amount of labeled intermediate training data, which can be too expensive to apply. Unlike existing DDTL algorithms, a novel feature-based [69] method benefits from multiple unlabeled source domains data with significant discrepancies. Furthermore, it can also handle multi-class classification and consistently produce promising results.

D. Frontier of TL Applications

In real-world problems, the most frequently and successfully applied ML technique is conventional supervised learning. After that, TL is predicted to be the next success in the industry. First of all, conventional ML algorithms cannot always meet the performance requirements due to the accuracy degradation caused by domain shifts. To address this issue, inductive TL [14], [88], [101]–[103] has started receiving more and more attention. Under the setting of inductive TL, multi-task learning is acknowledged as the most popular topic. Typically, it aims to improve the model robustness by using a small set of labeled target data set. Collecting a small set of labeled target data can decrease the training cost and enhance the robustness of the target model. The training process of inductive TL is the same as transductive TL. The only minor change is adding another target loss term to the final loss function of the model. However, the downside of inductive TL algorithms is that the training is more computationally expensive and time-consuming since another loss term is added.

Moreover, TL has been successfully applied to many applications in different fields, including but not limited to signal processing, sentiment analysis, health care system, and cyber-physical system (CPS).

Firstly, there are two main trends of signal processing, namely image processing [63], [88]–[91], audio analysis [92]–[94]. Transductive TL and distant domain TL are the main streams for this field. With TL algorithms, several different real-world problems can be solved by transferring knowledge from different domains with minimized cost. Sentiment analysis has also become an extremely active field in TL, including several applications: speech recognition, recommendation system, and spam detection. For example, the study [95] proposed the first TL enabled model for language understanding. A few works contributed a lot in cross-language translation [99], [100], [110] and sentiment analysis [96]–[98]. Furthermore, as more attention being brought to the health system, inductive and transductive TL has also been applied to solve many health cares and medical system-related problems, such as muscle fatigue classification [104], blood test analysis [101], [102], and medical imaging diagnosis [14], [88], [103]. Especially, TL methods also benefit a number of COVID-19 related problems [111]–[113], such as detection, treatment, and spread prediction.

Moreover, as a newly proposed concept, CPS requires moving beyond the classical fundamental computation and physics models. Therefore, it needs new models and theories that unify perspectives, capable of expressing the interacting dynamics and integration of a system's computational and physical components in a dynamic environment. A unified science would support composition, bridge the computational versus physical notions of time and space, cope with uncertainty, and enable

CPS to interoperate and evolve. Recently, there are many TL researches [105]–[108] conducted solid results in CPS.

VI. Open Challenges

So far, many studies of TL have carried out state-of-the results in several fields. Especially, transductive TL is the most active area in TL. However, there are still a number of open challenges of TL that are waiting to be addressed. This section discusses a number of major challenges in two levels: algorithm level and application level.

A. Challenges in Algorithms

We discuss several challenges at the algorithm level, such as human-guided TL, negative transfer, life-time TL, adversarial TL, and explainable TL.

First of all, most existing TL algorithms heavily rely on human instructions. Ideally, we expect models to learn an unseen task independently by using an algorithm to fully explore the data. The most successful case is AlphaZero [114] developed by Google Deepmind. It can teach itself how to master the Go game from scratch without any human experiences and instructions. However, the price of liberating the model is usually very high, and it requires a massive amount of time and computation power for the training. Therefore, the next direction is to lower the cost of this type of algorithms. In general, correctly inputting human pre-experience to the TL models can significantly reduce the time and the computation power required for training such a model. This concept is termed as human-guided TL. It aims to improve the efficiency of TL learning algorithms by correctly assembling human knowledge.

Secondly, negative transfer is widely acknowledged as an essential topic. It is one of the most significant limitations of TL. To address this issue, several distant domain TL algorithms [61], [63], [72] were proposed. Most existing methods are instance-based, and they are suffering from two major shortcomings: high case-dependence and massive source data requirement. Moreover, current methods can only transfer distant knowledge in different domains from the same modality. In other words, they can only transfer from image to image, audio to audio. Therefore, the next step of distant TL is to explore the potential of feature-based methods. Moreover, transferring knowledge between two different fields is one of the greatest challenges of distant TL, such as from image to audio and from text to image. Furthermore, an accurate domain distance measurement is also a critical factor in overcoming negative transfer. Commonly, MMD is the most popular non-parametric metric. However, it suffers from the risk of high-dimension data transformation. Other non-parametric metrics are not accurate enough for deep TL models. To address this issue, hybrid domain loss functions can help to improve the performance of distant TL.

The third, life-time TL is a relatively new concept. It aims to enable a TL framework with self-selecting the optimal learning method. The motivation behind this is that manually choosing a proper learning algorithm for a new task can be very time-consuming. Furthermore, we cannot do it manually when we are facing a new mission every time. Recently, a learning to transfer (L2T) framework [115] introduced a way to self-select an algorithm based on the input data. More importantly, there are not many studies regarding this issue. There is still a long way to go.

What is more, adversarial TL is becoming another focus in the field of TL. In general, it shares a similar idea to the original adversarial training pipeline. However, adversarial TL methods replace the feature generator with a distant feature extractor. There are a few proposed adversarial TL algorithms [7], but they are facing a critical difficulty in convergence. The convergence cannot be guaranteed in the training process due to the instability of the loss functions. Commonly, there are two counterparts in the final loss function, so the gradient explosion and disappear issues occur quite often. Therefore, designing stable loss functions will be the key to stabilize the training process for adversarial TL methods.

Furthermore, a high-level guideline for TL is also vital to the development of TL algorithms. When we develop a TL algorithm, a high-level guideline should provide comprehensive guidance to researchers in three main procedures of TL: 1) when to transfer, 2) what to transfer, and 3) how to transfer. Commonly, these three procedures can cover most high-level questions during the development of a TL algorithm. To the best of our knowledge, there are many guidance tools for conventional ML, but there is a lack of research for TL. A comprehensive guideline can help us develop algorithms and produce TL-based products in the industry.

B. Challenges in Applications

In TL applications, we demonstrate the current challenges into four major categories: Database for TL, Perception TL, User-machine Interaction, and Job Replacement. Moreover, these challenges are related to the algorithms, policy and ethics. Primarily, the database for TL focuses on data privacy, data labeling, data cleanness, and data sharing. Perception TL is mostly related to sentiment analysis for the applications that only take speech as the input. Moreover, user-machine interaction aims to develop more user-friendly products with TL techniques. Lastly, replacing job positions with TL-enabled machines are facing many ethics issues.

1) Challenges in Database for TL: First of all, the database is the cornerstone of all deep learning algorithms. The database has four main challenges: data privacy, data labeling, data cleanness, and data sharing. First, data privacy means that data sets cannot be shared due to restrictions, such as the patient information of medical data, copyrights of human face data, and security requirements of aviation data. Therefore, data sets with restricted information cannot be shared to the public. Moreover, some TL algorithms involve with multiple source data sets, so they have a greater chance of violating the rules and policies. To address this issue, an extra step to filter out classified information of data sets should be added to the process of data creation. What is more, many privacy-preserving methods have been adopted to supervised learning algorithms [116]. TL can also benefit from privacy-preserving techniques [117], [118]. However, this concept has not been well investigated due to the difficulties caused by multiple data sets in different domains. Importantly, it is critical to all the applications conducted by database companies and Internet-based products.

Secondly, data labeling is another issue in TL. Unlike traditional supervised learning, TL learning does not rely on a massive amount of labeled training data, so we do not need to manually label a big data set for TL. However, many deep TL learning models require multiple data sets in different domains but with the same label space. Therefore, it creates a new challenge of labeling data sets for TL, which requires to assign domain labels to multiple source domain data sets with the same instance label space. Moreover, it is relatively easy to discover several data sets with the same instance label space from different domains, but it is still time-consuming to manually assign domain labels when the source space is huge. In the future, creating exclusive data sets with domain labels can significantly benefit TL models. This problem is notably more critical to real-world applications with TL techniques because developing a real-world product requires way more data than academic experiments do.

2) Challenge in Perception TL Applications: Furthermore, perception TL concentrates on the verbal and motional inputs taken by TL algorithms, such as speech, voice, and motions. Recently, the stationary image data is considered as the most common input of most ML-enabled applications, such as auto-driving systems, smart wearings, and security systems. The easiest access is the reason why the majority of the ML-based applications most prefer the stationary image data. However, there are four major drawbacks of the stationary image input. Firstly, most existing applications are not friendly to people with disabilities. For example, stationary image-based products can cause difficulties for people who cannot type the keyboard due to their disabilities. Secondly, stationary image-based ML systems cannot easily be controlled by users. The third, ML-enabled security systems with image-based inputs suffer from safety issues because image data can be easily faked. Lastly, the domain shift can hurt the performance significantly.

To address these issues, many studies proposed to adopt other data types with less accessibility can for a wide range of applications by adopting TL techniques. For example, the study [95] proposed the first TL algorithm for speech recognition and achieved a promising performance. Furthermore, TL algorithms have been expanded to other areas: gesture recognition, voice recognition, and Micro-expression recognition. Therefore, the next stage of TL-based applications is to expand the types of input sources and enable multiple types of input sources. However, there are many unsolved problems in TL models for other types of inputs. The most challenging topic is sentiment input, such as speech and text. For example, most products can only take key-words as inputs but cannot handle longer sentences. There are many successful TL algorithms for image processing tasks, but there are not many studies in sentiment analysis. Recently, some works [96], [119] have introduced TL algorithms for sentiment-focused long speech analysis. Therefore, adapting TL techniques to real-world products with perception inputs is a very challenging topic.

VII. Conclusion

Finally, the number of TL-related researches has been on a rapid increase in the past decade. Moreover, its usage in industries is bypassing supervised learning due to its advantages on efficiency and performance. In the future, with the above four main challenges being addressed, TL will be more widely used in both academia and industry.

References

[1] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[2] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[3] F. M. Cariucci, L. Porzi, B. Caputo, E. Ricci, and S. R. Bul, "Autodial: Automatic domain alignment layers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5077–5085.

[4] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, p. 9, 2016.

[5] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," 2019, *arXiv:1911.02685*.

[6] O. Day and T. M. Khoshgoftaar, "A survey on heterogeneous transfer learning," *J. Big Data*, vol. 4, no. 1, p. 29, 2017.

[7] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "Survey on deep transfer learning," in *Proc. Int. Conf. Artif. Neural Netw.* Springer, 2018, pp. 270–279.

[8] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community," *J. Appl. Remote Sens.*, vol. 11, no. 4, 2017, Art. no. 042609.

[9] A. N. Soni, "Application and analysis of transfer learning-survey," *Int. J. Sci. Res. Eng. Develop.*, vol. 1, no. 2, pp. 272–278, 2018.

[10] R. Liu, Y. Shi, C. Ji, and M. Jia, "A survey of sentiment analysis based on transfer learning," *IEEE Access*, vol. 7, pp. 85401–85412, 2019.

[11] W. Pan, "A survey of transfer learning for collaborative recommendation with auxiliary data," *Neurocomputing*, vol. 177, pp. 447–453, 2016.

[12] A. Sufian, A. Ghosh, A. S. Sadiq, and F. Smarandache, "A survey on deep transfer learning to edge computing for mitigating the covid-19 pandemic," *J. Syst. Architecture*, vol. 108, 2020, Art. no. 101830.

[13] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: A survey," *Knowl.-Based Syst.*, vol. 80, pp. 14–23, 2015.

[14] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280–296, 2019.

[15] Q. Wu *et al.*, "Online transfer learning with multiple homogeneous or heterogeneous sources," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 7, pp. 1494–1507, Jul. 2017.

[16] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.

[17] M. Ghifary, W. B. Kleijn, and M. Zhang, "Domain adaptive neural networks for object recognition," in *Proc. Pacific Rim Int. Conf. Artif. Intell*, Springer, 2014, pp. 898–904.

[18] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3071–3085, Dec. 2019.

[19] R. Xia, C. Zong, X. Hu, and E. Cambria, "Feature ensemble plus sample selection: Domain adaptation for sentiment classification," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 10–18, Jun. 2013.

[20] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7167–7176.

[21] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, *ser. ICML15. JMLR.org*, 2015, pp. 1180–1189. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045118.3045244

[22] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 801–814, Apr. 2019.

[23] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–450.

[24] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2272–2281.

[25] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 136–144.

[26] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5018–5027.

[27] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70. JMLR.org, 2017, pp. 2208–2217.

[28] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, "Adaptive batch normalization for practical domain adaptation," *Pattern Recognit.*, vol. 80, pp. 109–117, 2018.

[29] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4068–4076.

[30] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[31] R. Xia, X. Hu, J. Lu, J. Yang, and C. Zong, "Instance selection and instance weighting for cross-domain sentiment classification via PU learning," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2176–2182.

[32] F. Xu, J. Yu, and R. Xia, "Instance-based domain adaptation via multiclustering logistic approximation," *IEEE Intell. Syst.*, vol. 33, no. 1, pp. 78–88, Jan./Feb. 2018.

[33] R. Xia, J. Yu, F. Xu, and S. Wang, "Instance-based domain adaptation in nlp via in-target-domain logistic approximation," in *Proc. 28th AAAI Conf. Artif. Intell.*, vol. 28, no. 1, 2014.

[34] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1855–1862.

[35] H. Daumé III, "Frustratingly easy domain adaptation," 2009, *arXiv:0907.1815*.

[36] D. S. Maitra, U. Bhattacharya, and S. K. Parui, "CNN based common approach to handwritten character recognition of multiple scripts," in *Proc. 13th Int. Conf. Document Anal. Recognit.*, 2015, pp. 1021–1025.

[37] M. Fang, Y. Guo, X. Zhang, and X. Li, "Multi-source transfer learning based on label shared subspace," *Pattern Recognit. Lett.*, vol. 51, pp. 101–106, 2015.

[38] M. J. Afridi, A. Ross, and E. M. Shapiro, "On automated source selection for transfer learning in convolutional neural networks," *Pattern Recognit.*, vol. 73, pp. 65–75, 2018.

[39] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[40] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[41] S. Niu, J. Wang, Y. Liu, and H. Song, "Transfer learning based data-efficient machine learning enabled classification," in *Proc. IEEE Intl Conf. Dependable, Autonomic Secure Comput., Intl Conf. Pervasive Intell. Comput., Intl Conf Cloud Big Data Comput., Intl Conf. Cyber Sci. Technol. Congr.*, 2020, pp. 620–626.

[42] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proc. 24th Int. Conf. Mach. Learn., ser.*, 2007, pp. 193–200.

[43] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *J.-Japanese Soc. Artif. Intell.*, vol. 14, no. 771–780, p. 1612, 1999.

[44] G. Matasci, D. Tuia, and M. Kanevski, "SVM-based boosting of active learning strategies for efficient domain adaptation," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 5, no. 5, pp. 1335–1343, Oct. 2012.

[45] Z. Li, B. Liu, and Y. Xiao, "Cluster and dynamic-tradaboost-based transfer learning for text classification," in *Proc. 13th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discov.*, Jul. 2017, pp. 2291–2295.

[46] Z. Yuan, D. Bao, Z. Chen, and M. Liu, "Integrated transfer learning algorithm using multi-source tradaboost for unbalanced samples classification," in *Proc. Int. Conf. Comput. Intell. Inf. Syst.*, Apr. 2017, pp. 188–195.

[47] S. Al-Stouhi and C. K. Reddy, "Adaptive boosting for transfer learning using dynamic updates," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discov. Databases - Volume Part I*, 2011, pp. 60–75. [Online]. Available: http://dl.acm.org/citation.cfm?id=2034063.2034080

[48] D. Pardoe and P. Stone, "Boosting for regression transfer," in *Proc. 27th Int. Conf. Int. Conf. Mach. Learn.*, 2010, pp. 863–870. [Online]. Available: http://dl.acm.org/citation.cfm?id=3104322.3104432

[49] Z. Zhao, Y. Chen, J. Liu, Z. Shen, and M. Liu, "Cross-people mobile-phone based activity recognition," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011.

[50] M. Long, Z. Cao, J. Wang, and S. Y. Philip, "Learning multiple tasks with multilinear relationship networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1594–1603.

[51] W.-Y. Deng, Q.-H. Zheng, and Z.-M. Wang, "Cross-person activity recognition using reduced kernel extreme learning machine," *Neural Netw.*, vol. 53, pp. 1–7, 2014.

[52] H. Li, Y. Shi, Y. Liu, A. G. Hauptmann, and Z. Xiong, "Cross-domain video concept detection: A. joint discriminative and generative active learning approach," *Expert Syst. Appl.*, vol. 39, no. 15, pp. 12 220–12 228, 2012.

[53] F. Nater, T. Tommasi, H. Grabner, L. Van Gool, and B. Caputo, "Transferring activities: Updating human behavior analysis," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 1737–1744.

[54] Z. Wang, Y. Song, and C. Zhang, "Transferred dimensionality reduction," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2008, pp. 550–565.

[55] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Self-taught clustering," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 200–207.

[56] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 759–766. [Online]. Available: http://doi.acm.org/10.1145/1273496.1273592

[57] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "Youtube-Boundingboxes: A large high-precision human-annotated data set for object detection in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7464–7473.

[58] L. Duan, D. Xu, and S. Chang, "Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1338–1345.

[59] L. Ge, J. Gao, H. Ngo, K. Li, and A. Zhang, "On handling negative transfer and imbalanced distributions in multiple source transfer learning," *Stat. Anal. Data Mining: The ASA Data Sci. J.*, vol. 7, no. 4, pp. 254–271, 2014.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[61] B. Tan, Y. Song, E. Zhong, and Q. Yang, "Transitive transfer learning," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2015, pp. 1155–1164.

[62] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," in *Proc. 30th Assoc. Adv. Artif. Intell. Conf. Artif. Intell.*, vol. 30, no. 1, 2016.

[63] B. Tan, Y. Zhang, S. J. Pan, and Q. Yang, "Distant domain transfer learning," in *Proc. 31th AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017.

[64] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu, "Translated learning: Transfer learning across different feature spaces," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 353–360.

[65] G.-J. Qi, C. Aggarwal, and T. Huang, "Towards semantic knowledge propagation from text corpus to web images," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 297–306.

[66] Y. Zhu *et al.*, "Heterogeneous transfer learning for image classification," *Assoc. Adv. Artif. Intell.*, vol. 11, pp. 1304–1309, 2011.

[67] Q. Wu, M. K. Ng, and Y. Ye, "Cotransfer learning using coupled markov chains with restart," *IEEE Intell. Syst.*, vol. 29, no. 4, pp. 26–33, Jul./Aug. 2014.

[68] M. K. Ng, Q. Wu, and Y. Ye, "Co-transfer learning via joint transition probability graph based method," in *Proc. 1st Int. Workshop Cross Domain Knowl. Discov. Web Social Netw. Mining*, 2012, pp. 1–9.

[69] S. Niu, H. Yihao, J. Wang, Y. Liu, and H. Song, "Feature-based distant domain transfer learning," in *Proc. IEEE Int. Conf. Big Data*, 2020.

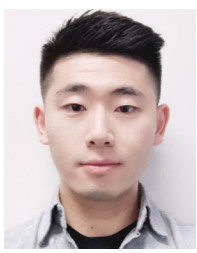
[70] J. Lee, P. Sattigeri, and G. W. Wornell, "Learning new tricks from old dogs: Multi-source transfer learning from pre-trained networks," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., vol. 32, 2019, pp. 4370–4380, [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/6048ff4e8cb07aa60b6777b6f7384d52-Paper.pdf

[71] C. Bell, "Mutual information and maximal correlation as measures of dependence," *Ann. Math. Statist.*, vol. 33, no. 2, pp. 587–595, 1962.

[72] B. Wang, J. Mendez, M. Cai, and E. Eaton, "Transfer learning via minimizing the performance gap between domains," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 10 644–10 654.

[73] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2006, pp. 120–128.

[74] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[75] Y. Yao and G. Doretto, "Boosting for transfer learning with multiple sources," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1855–1862.
[76] E. Eaton *et al.*, "Selective transfer between learning tasks using task-based boosting," in *Proc. 25th AAAI Conf. Artif. Intell.*, AAAI'11, San Francisco, California: AAAI Press, 2011, pp. 337–342.
[77] J. Jiang and C. Zhai, "Instance weighting for domain adaptation in NLP," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics*, 2007, pp. 264–271.
[78] X. Liao, Y. Xue, and L. Carin, "Logistic regression with an auxiliary data source," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 505–512. [Online]. Available: http://doi.acm.org/10.1145/1102351.1102415
[79] Y. Wu, W. Li, M. Minoh, and M. Mukunoki, "Can feature-based inductive transfer learning help person re-identification?," in *Proc. IEEE Int. Conf. Image Process.*, 2013, pp. 2812–2816.
[80] L. Mihalkova, T. Huynh, and R. J. Mooney, "Mapping and revising markov logic networks for transfer learning," *Assoc. Adv. Artif. Intell.*, vol. 7, 2007, pp. 608–614.
[81] L. Mihalkova and R. J. Mooney, "Transfer learning by mapping with minimal target data," in *IJCAI*, 2009, pp. 1163–1168, [Online]. Available: http://ijcai.org/Proceedings/09/Papers/196.pdf
[82] J. Davis and P. Domingos, "Deep transfer via second-order markov logic," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 217–224.
[83] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*.
[84] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
[85] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[86] R. T. Ng and J. Han, "E cient and E ective clustering methods for spatial data mining," in *Proc. IEEE Symp. Very Large-Scale Data Anal. Vis.*, 1994, pp. 144–155.
[87] Z. Wang, Z. Dai, B. Póczos, and J. Carbonell, "Characterizing and avoiding negative transfer," *CoRR*, vol. abs/1811.09751, 2019, *arXiv:1811.09751*, [Online]. Available: http://arxiv.org/abs/1811.09751
[88] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
[89] Y. Yuan, X. Zheng, and X. Lu, "Hyperspectral image superresolution by transfer learning," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1963–1974, May. 2017.
[90] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, and A. Agrawal, "Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection," *Construction Building Mater.*, vol. 157, pp. 322–330, 2017.
[91] M. M. Ghazi, B. Yanikoglu, and E. Aptoula, "Plant identification using deep neural networks via optimization of transfer learning parameters," *Neurocomputing*, vol. 235, pp. 228–235, 2017.
[92] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," 2017, *arXiv:1703.09179*.
[93] Y. Jia *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4480–4490.
[94] A. Kumar, M. Khadkevich, and C. Fgen, "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2018, pp. 326–330.
[95] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[96] X. L. Dong and G. De Melo, "A helping hand: Transfer learning for deep sentiment analysis," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics* (Volume 1: Long Papers), 2018, pp. 2524–2534.
[97] J. Islam and Y. Zhang, "Visual sentiment analysis for social images using transfer learning approach," in *Proc. IEEE Int. Conf. Big Data Cloud Comput., Social Comput. Netw., Sustain. Comput. Commun.*, 2016, pp. 124–130.
[98] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," 2018, *arXiv:1801.06353*.
[99] J. T. Zhou, S. J. Pan, I. W. Tsang, and S.-S. Ho, "Transfer learning for cross-language text categorization through active correspondences construction," in *Proc. 30th AAAI Conf. Artif. Intell.*, AAAI'16, Phoenix, Arizona: AAAI Press, 2016, pp. 2400–2406.
[100] G. Pham, D. Donovan, Q. Dam, and A. Contant, "Learning words and definitions in two languages: What promotes cross-language transfer?" *Lang. Learn.*, vol. 68, no. 1, pp. 206–233, 2018.
[101] L. H. Vogado, R. M. Veras, F. H. Araujo, R. R. Silva, and K. R. Aires, "Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification," *Eng. Appl. Artif. Intell.*, vol. 72, pp. 415–422, 2018.
[102] R. Colbaugh, K. Glass, and G. Gallegos, "Ensemble transfer learning for alzheimer's disease diagnosis," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2017, pp. 3102–3105.
[103] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *J. Med. Imag.*, vol. 3, no. 3, 2016, Art. no. 034501 .
[104] N. Siekirk, Q. Lai, and B. Kendall, "Effects of limb-specific fatigue on motor learning during an upper extremity proprioceptive task," *Int. J. Motor Control Learn.*, vol. 1, no. 1, pp. 76–81, 2018.
[105] J. Lee, M. Azamfar, J. Singh, and S. Siahpour, "Integration of digital twin and deep learning in cyber-physical systems: Towards smart manufacturing," *IET Collaborative Intell. Manuf.*, vol. 2, no. 1, pp. 34–36, 2020.
[106] T. Hou, G. Feng, S. Qin, and W. Jiang, "Proactive content caching by exploiting transfer learning for mobile edge computing," *Int. J. Commun. Syst.*, vol. 31, no. 11, 2018, Art. no. e3706.
[107] S. Loidl, "Towards pervasive learning: Welearn. Mobile. A CPS package viewer for handhelds," *J. Netw. Comput. Appl.*, vol. 29, no. 4, pp. 277–293, 2006.
[108] H. Zou, Y. Zhou, H. Jiang, B. Huang, L. Xie, and C. Spanos, "Adaptive localization in dynamic indoor environments by transfer kernel learning," in *Proc. IEEE Wirel. Commun. Netw. Conf.*, 2017, pp. 1–6.
[109] W. Zhang *et al.*, "Deep model based transfer and multi-task learning for biological image analysis," IEEE Trans. Big Data, vol. 6, no. 2, pp. 322–333, Jun. 2020.
[110] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Cross-language transfer learning for deep neural network based speech enhancement," in *Proc. 9th Int. Symp. Chinese Spoken Lang. Process.*, 2014, pp. 336–340.
[111] Y. Pathak, P. K. Shukla, A. Tiwari, S. Stalin, S. Singh, and P. K. Shukla, "Deep transfer learning based classification model for covid-19 disease," IRBM, 2020, doi: 10.1016/j.irbm.2020.05.003.
[112] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: Automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, Springer, vol. 43, no. 2, pp. 635–640, 2020.
[113] M. Loey, F. Smarandache, and N. E. M. Khalifa, "Within the lack of chest covid-19 x-ray dataset: A novel detection model based on gan and deep transfer learning," *Symmetry*, vol. 12, no. 4, p. 651, 2020.
[114] D. Silver *et al.*, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.
[115] W. Ying, Y. Zhang, J. Huang, and Q. Yang, "Transfer learning via learning to transfer," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5085–5094.
[116] P. Li *et al.*, "Multi-key privacy-preserving deep learning in cloud computing," *Future Gener. Comput. Syst.*, vol. 74, pp. 76–85, 2017.
[117] D. Gao, Y. Liu, A. Huang, C. Ju, H. Yu, and Q. Yang, "Privacy-preserving heterogeneous federated transfer learning," in *Proc. IEEE Int. Conf. Big Data*, 2019, pp. 2552–2559.
[118] Q. Li, Z. Wen, and B. He, "Federated learning systems: Vision, hype and reality for data privacy and protection," 2019, *arXiv:1907.09693*.
[119] S. Zhang, X. Zhang, J. Chan, and P. Rosso, "Irony detection via sentiment-based transfer learning," *Inf. Process. Manage.*, vol. 56, no. 5, pp. 1633–1644, 2019.



Shuteng Niu is currently working toward the Ph.D. degree with the Department of Electrical Engineering and Computer Science, Embry-Riddle Aeronautical University, Daytona Beach, FL, USA. His main research interests include machine learning, data mining, and medical imaging.

Yongxin Liu received the first Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2018. He is currently working toward the second Ph.D. degree with the Department of Electrical Engineering and Computer Science, Embry-Riddle Aeronautical University, Daytona Beach, FL, USA. His main research interests include data mining, wireless networks, the Internet of Things, and unmanned aerial vehicles.

Jian Wang received the B.S. degree from Nanyang Normal University, Nanyang, China, in 2014 and the M.S. degree from South China Agricultural University, Guangzhou, China, in 2017. He is currently working toward the Ph.D. degree with the Department of Electrical Engineering and Computer Science, Embry-Riddle Aeronautical University, Daytona Beach, FL, USA. His main research interests include wireless networks, unmanned aerial systems, and machine learning.

Houbing Song (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, USA, in 2012. In August 2017, he joined the Department of Electrical Engineering and Computer Science, Embry-Riddle Aeronautical University, Daytona Beach, FL, USA, where he is currently an Assistant Professor and the Director with the Security and Optimization for Networked Globe Laboratory (SONG Lab). He is the author of more than 100 articles. His research interests include AI or machine learning, big data analytics, cyber-physical systems, cybersecurity and privacy, Internet of Things, and unmanned aircraft systems. Dr. Song has been an Associate Technical Editor for the *IEEE Communications Magazine* since 2017, an Associate Editor for the IEEE INTERNET OF THINGS JOURNAL since 2020 and the IEEE JOURNAL ON MINIATURIZATION FOR AIR AND SPACE SYSTEMS since 2020, and the Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the IEEE INTERNET OF THINGS JOURNAL, the IEEE NETWORK, the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the IEEE SENSORS JOURNAL, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, and the IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS. He is the Editor of six books, including *Big Data Analytics for Cyber-Physical Systems: Machine Learning for the Internet of Things* (Elsevier, 2019), *Smart Cities: Foundations, Principles and Applications* (Wiley, 2017), *Security and Privacy in Cyber-Physical Systems: Foundations, Principles and Applications* (Wiley-IEEE Press, 2017), *Cyber-Physical Systems: Foundations, Principles and Applications* (Academic Press, 2016), and *Industrial Internet of Things: Cybermanufacturing Systems* (Springer, 2016). His research has been featured by popular news media outlets, including the IEEE GlobalSpec's Engineering360, the Association for Unmanned Vehicle Systems International (AUVSI), the Fox News, the USA Today, the U.S. News & World Report, the Forbes, The Washington Times, the WFTV, and the New Atlas. He is a Senior Member of ACM and an ACM Distinguished Speaker. He was the recipient of the Best Paper Award from the 12th IEEE International Conference on Cyber, Physical and Social Computing (CPSCom-2019), the Best Paper Award from the 2nd IEEE International Conference on Industrial Internet (ICII 2019), the Best Paper Award from the 19th Integrated Communication, Navigation and Surveillance technologies (ICNS 2019) Conference, the Best Paper Award from the 6th IEEE International Conference on Cloud and Big Data Computing (CBDCom 2020), and the Best Paper Award from the 15th International Conference on Wireless Algorithms, Systems, and Applications (WASA 2020).