Transfer Adaptation Learning: A Decade Survey

Lei Zhang, *Senior Member, IEEE*

Abstract—The world we see is ever-changing and it always changes with people, things, and the environment. *Domain* is referred to as the state of the world at a certain moment. A research problem is characterized as *domain transfer adaptation* when it needs knowledge correspondence between different moments. Conventional machine learning aims to find a model with the minimum expected risk on test data by minimizing the regularized empirical risk on the training data, which, however, supposes that the training and test data share similar joint probability distribution. *Transfer adaptation learning* aims to build models that can perform tasks of target domain by learning knowledge from a semantic related but distribution different source domain. It is an energetic research filed of increasing influence and importance. This paper surveys the recent advances in transfer adaptation learning methodology and potential benchmarks. Broader challenges being faced by transfer adaptation learning researchers are identified, i.e., instance re-weighting adaptation, feature adaptation, classifier adaptation, deep network adaptation, and adversarial adaptation, which are beyond the early semi-supervised and unsupervised split. The survey provides researchers a framework for better understanding and identifying the research status, challenges and future directions of the field.

Index Terms—Transfer Learning, Domain Adaptation, Distribution Discrepancy, Computer Vision

1 Introduction

VISUAL understanding of an image or video is a long-standing and challenging problem in computer vision. Visual classification, as a fundamental problem of visual understanding, aims to recognize *what* an image depicts. A solidified route of visual classification is to establish a learning model by collecting an image dataset, which can be recognized as *target* data. However, labeling a large number of target samples is cost-ineffective which consumes a lot of human resources in labor and time expenses and becomes almost unrealistic. Therefore, leveraging another distribution different but semantic related *source* domain with sufficiently labeled samples for recognizing task samples is becoming an increasingly important topic.

With the explosive increase of multi-source data from Internet such as YouTube and Flickr, a large number of labeled web database can be easily crawled. It is thus natural to consider train a learning model using multi-source web data for recognizing target data. However, a prevailing problem is that distribution mismatch and domain shift [1], [2] across source and target domain often exist owing to various factors such as resolution, illumination, viewpoint, background, etc. in computer vision. Therefore, the classification performance is dramatically degraded when the source data that used to learn the classifier model has different distribution from the target data on which the model is applied. This is due to that the fundamental independent identical distribution condition supposed in statistical learning is no longer satisfied, which, therefore promotes the emergence of transfer learning (TL) and domain adaptation (DA) [3], [4], [5]. In early, TL supposed different joint probability distribution, i.e., $P(X_{source}, Y_{source}) \neq P(X_{target}, Y_{target})$ between source and target domains. DA supposed different marginal distribution, i.e., $P(X_{source}) \neq P(X_{target})$ but similar category space between domains i.e., $P(Y_{source}|X_{source}) = P(Y_{target}|X_{target})$. Several related reviews on transfer learning and domain adaptation can be referred to as [4], [6], [7], [8], [9], [10], [11], [12], [13]. In this paper, we use a general name *Transfer Adaptation Learning (TAL)* for unifying both TLs and DAs. In the past decade, TAL was an active area in machine learning community, and the goal of which is to narrow down the distribution gap between source and target data, such that the labeled source data from one or more relevant domains can be utilized for executing tasks in target domain, as illustrated in Fig. 1.

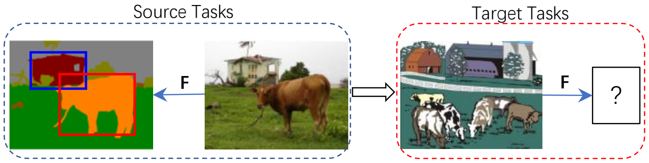

Fig. 1. Cross-domain object detection, recognition and semantic segmentation. $\mathbf{F}$ denotes models of three tasks learned on source domain.

Moving forward, deep learning (DL) techniques [14], [15], [16], [17] have recently become dominant and powerful algorithms in feature representation and abstraction for image classification. In particular, the parameter adaptability and generality of DL models to other target data is worthy of praise, by fine-tuning a pre-trained deep neural network using a small amount of target data. Therefore, *fine-tune* has become a commonly used strategy for training deep models and frameworks in various applications, such as object detection [18], [19], [20], [21], person re-identification [22], [23], [24], medical imaging [25], [26], [27], remote sensing [28], [29], [30], [31], etc. Generally, the *fine-tune* can be recognized as a prototype for bridging the big source data and the target data [32], which also facilitates the research of visual transfer learning in computer vision. Conceptually, *fine-tune* is big

- *L. Zhang is with the School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China. (E-mail: leizhang@cqu.edu.cn).*

data-driven transfer learning method, which depends on a pre-trained model with a big source database. The context of transfer learning challenge and why pre-training of representations can be useful have been formally explored in [32]. Extensively, from the viewpoint of generative learning, the popular generative adversarial net (GAN) [33] and its variants [34], [35], [36], [37], [38] that aim to synthesize plausible images of some target distribution from, for example, the noise signal (source distribution), can also be recognized as generalized transfer learning techniques. Differently, conventional transfer learning approaches put emphasis on the *output* knowledge (high-level model parameters) adaptation across source and target domains, while GANs focus on *input* data (low-level image pixels) adaptation from source distribution to target distribution. Recently, image pixel-level transfer has been intensively studied in image-to-image translation [39], [40], [41], [42], [43], style transfer [44], [45], [46] and target face synthesis (e.g., pose transfer vs. age transfer) [47], [48], [49], [50], [51], etc.

In this paper, we focus on technical advances and challenges in model-driven transfer adaptation learning. Learning from multiple sources for transferring or adapting to new target domains offers the possibility of promoting model generalization and understanding the biological learning essence. Transfer adaptation learning is similar but different from multi-task learning that resorts to the shared feature representations or classifiers for related tasks [52], simultaneously. In the past decade, a number of transfer learning and domain adaptation approaches have been emerged. In this paper, the challenges and advances in the research field of transfer learning and domain adaptation are identified and surveyed. Specifically, we explore five key challenges of transfer adaptation learning, which are beyond the early semi-supervised and unsupervised split.

- *Instance Re-weighting Adaptation*. Due to the probability distribution discrepancy across domains, it is natural to account for the difference by directly inferring the resampling weights of instances based on feature distribution matching across source and target data in a non-parametric manner. The parameter estimation of the weights under a parametric distribution assumption remains to be a challenge.
- *Feature Adaptation*. For adapting the data from multiple sources, learning a common feature subspace or representation where the projected source and target domain are with similar distribution is generally resulted. The heterogeneity of data distribution makes it challenging to gain such generic feature.
- *Classifier Adaptation*. The classifier trained on instances of source domain is often biased when recognizing instances from target domain due to the domain shifts. Learning a generalized classifier from multiple domains that can be used for other different domains, is a challenging topic.
- *Deep Network Adaptation*. Deep neural networks have been recognized with strong feature representation power and general deep model is built on single domain. Large domain shift makes deep neural network training challenging to obtain transferrable deep representation.
- *Adversarial Adaptation*. Adversarial learning originates from the generative adversarial nets. The objective of TL/DA is to make the source and target domains more close in feature space. It is amount to confusing the two domains, such that they can not be easily discriminated. Therefore, there comes a technical challenge in domain confusion by using adversarial training and gaming strategy.

For each challenge, the taxonomic classes and sub-classes are presented to structure the recent work in transfer adaptation learning. We start with an discussion of weakly-supervised learning perspectives in Section 2, which is followed by the technical advances in transfer adaptation learning, including instance re-weighting adaptation (Section 3), feature adaptation (Section 4), classifier adaptation (Section 5), deep network adaptation (Section 6), and adversarial adaptation (Section 7). The existing benchmarks and future challenging tasks are discussed in Section 8 and the paper is concluded in Section 9.

2 Weakly-supervised Learning Perspective

The concept of *weak learning* originated 20 years ago in AdaBoost [53] and Ensemble learning [54] algorithms, which tend to ensemble multiple weak learners to solve a problem. AdaBoost, that has been listed as the top 10 algorithms in data mining [55], aims to learn multiple weak learners, in which each weak learner is obtained by training on the weighted incorrectly classified examples. By ensemble of multiple weak learners, the performance is significantly boosted. Although the *weak* concept was proposed as early as 1997, the problem in that era was still established on strong supervision due to the relatively *smaller* data. That is, the early problem can be strongly learned by conventional statistical learning models. However, today, the big data era, the problem becomes really a weak supervision problem, due to the *inaccurate*, *inexact*, and *incomplete* characteristics of data labels [56], which, therefore, has to be weakly learned. Currently, weakly-supervised learning is becoming a leading research topic. Undoubtedly, transfer adaptation learning, that resorts to solving cross-domain problems, is also a kind of weakly-supervised learning methodology. This section is deployed with typical weakly-supervised learning frameworks and perspectives.

2.1 Semi-supervised Learning

Semi-supervised learning (SSL) aims to solve the problem where there are a large amount of unlabeled examples X_u and a few labeled examples (X_l, Y_l) in the dataset [57], [58]. Generally, semi-supervised learning methods consist of four categories. (i) Generative methods that advocate generating the labeled and unlabeled data via an inherent model [59], [60], [61]. (ii) Low-density separation methods that constrains the classifier boundary crossing the low-density region [62], [63], [64]. (iii) Disagreement based methods that advocate co-training of multiple learners for annotating the unlabeled instances [65], [66], [67]. (iv) Graph based methods that propose to build the connection graph of the training instances for label propagation through graph modeling [68], [69], [70], [71]. A good literature review of semi-supervised learning can be referred to as [72], [73].

Consider a general SSL framework, then the following expected risk is generally minimized.

$$R[P_r, W, l(X, Y, W)] = E_{(X,Y)\sim P_r}[l(X, Y, W)] \qquad (1)$$

where P_r is the probability distribution, $X = [X_l, X_u] \in \Re^{D\times N}$ is the data, $Y = [Y_l, 0] \in \Re^{C\times N}$ is the label index in which zero vector is posed for unlabeled samples, W is the model parameter. D, N and C denote the number of dimensionality, samples, and classes of data, respectively.

The training data usually comes from a subset, therefore, the regularized risk, i.e., the average empirical risk with regularization is minimized.

$$R_{reg}[W, l(X, Y, W)] = R_{emp}[W, l(X, Y, W)] + \lambda\Omega(W) \quad (2)$$

where $l(\cdot)$ is the prediction loss function and $R_{emp}[\cdot]$ is the average empirical risk (e.g. mean squared loss) on training data. A general SSL model with graph based manifold regularization can be written as

$$\min_{W} R_{reg}[W, l(X, Y, W)] + \gamma \sum_{i,j}^{N} A_{i,j} d^2(f_i, f_j) \qquad (3)$$

where f_i is the predicted label for sample i and A is the affinity matrix used for locality preservation. Usually, $A_{i,j} = \exp(-\sigma d^2(x_i, x_j))$ if x_i and x_j are neighbors, otherwise 0.

The key difference from transfer learning is that the marginal distribution and label space distribution are the same, i.e., $P(X_l) = P(X_u)$ and $P(Y_l|X_l) = P(Y_u|X_u)$. Generally, SSL attempts to exploit the unlabeled data for auxiliary training on the labeled data without human intervention, because the distribution of unlabeled data can intrinsically reflect sample class information. Actually, in SSL model, three basic assumptions, i.e., *smoothness*, *cluster*, and *manifold*, have been established. The *smoothness* assumption denotes that data is distributed with different density, and the two instances falling into the region of high density have the same label. The *cluster* assumption denotes that data have inherent cluster structure, and the two samples in the same cluster are more similar. The *manifold* assumption means that the data lie on a manifold, and the instances in a small local neighborhood have similar semantics. The three basic assumptions are visually shown in Fig. 2.

2.2 Active Learning

Active learning (AL) aims to obtain the ground-truth labels of selected unlabeled instances with human intervention [74], [75], which is different from semi-supervised learning that exploits unlabeled data together with labeled data for improving recognition performance. Specifically, AL aims at progressively selecting and annotating the most informative data points from the pool of unlabeled samples, such that the labeling cost for training an effective model can be minimized [76], [77]. There are two engines, learning engine and selection engine, in active learning paradigm. The learning engine targets at obtaining a baseline classifier, while the selection engine tries to select unlabeled instances and deliver them to human experts for manual annotation. The selection criteria is generally determined based on information uncertainty [74], [75].

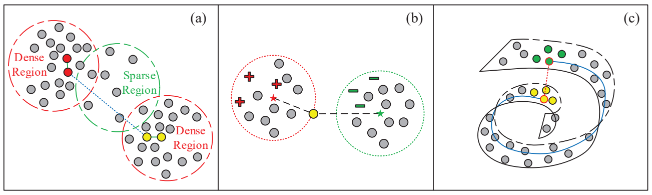

Fig. 2. Illustration of the three basic assumptions in SSL. a) *Smoothness* assumption. b) *Cluster* assumption. c) *Manifold* assumption.

2.3 Zero-shot Learning

Recently, zero-shot learning (ZSL) [78], [79], [80], [81], [82], as a typical weakly-supervised learning paradigm, has attracted researchers' attention. ZSL tries to recognize the samples of unseen categories that never appear in training data, i.e., there is no overlap between the seen categories in training data and the unseen categories in test data. That is, the label space distribution between training and test data is different, i.e., $P(Y_{seen}|X_{seen}) \neq P(Y_{unseen}|X_{unseen})$, which can be recognized as a special case of transfer learning. This situation often occurs in various fields, due to that manually annotating tens of thousands of different object classes in the world is quite expensive and almost unrealistic. The general problem of ZSL is as follows.

Zero-shot learning with disjoint training and testing classes. *Let $\mathcal{X}$ be an arbitrary feature space of training data. Let $\mathcal{Y}$ and $\mathcal{Z}$ be the sets of seen and unseen object categories, respectively, and there is $\mathcal{Y} \bigcap \mathcal{Z} = \emptyset$. The task is to learning a classifier f: $\mathcal{X} \mapsto \mathcal{Z}$ by using the training data $(x_1, y_1), \cdots, (x_N, y_N) \subset (\mathcal{X}, \mathcal{Y})$.*

An extension of ZSL is the one/few shot learning (O/FSL) where few labeled examples of each unseen object classes are revealed during training process. The usual idea of Z/O/FSL is to learn the embedding of the image feature into the semantic space or semantic attributes [79], [83]. Afterwards, recognition of new classes can be conducted by matching the semantic embedding of the visual features with the semantic/attribute representation. However, visual-semantic mapping learned from the seen categories may not generalize well to the unseen category due to the domain shift, which, thus can be a challenging topic by utilizing transfer learning to ZSL. Actually, for improving ZSL under domain shifts, transductive or semi-supervised zero-shot learning approaches have been studied for reducing the difference of visual-semantic mappings between seen and unseen categories [84], [85], [86], [87], [88].

2.4 Open Set Recognition

Conventional recognition tasks in computer vision where all testing classes are known at training time are generally recognized as *closed-set* recognition. Open set recognition addresses a more realistic vision scenario where unknown classes can be encountered during testing time [89], [90], [91], [92], which shares very similar characteristic with ZSL in tasks. ZSL is different from open set recognition that the former uses the semantic embedding of visual features for recognizing unknown classes, while the latter focus on a one-class classification problem.

More recently, a similar open set framework with transductive ZSL for recognition under domain shift is the *open-*

set domain adaptation approach [93], [94], which were established on the concept of open set recognition. Conventional domain adaptation assumes that the categories in target domain are known and can be seen in the source domain, while open-set domain adaptation addresses the scenarios where the target domain contains the instances of categories that are unseen in the source domain [94]. The differences between zero-shot learning and open-set domain adaptation lie in that (1) ZSL tends to solve the recognition of instances of unseen categories under the *same* marginal distribution across training and testing data, while open set domain adaptation aims to solve the same problem but under *different* marginal distribution across source and target domains. (2) *Generalized* ZSL [83], [95] was proposed for the scenario where the training and test classes are not necessarily disjoint, while open set domain adaptation was proposed for the scenario where there still a few categories of interest are shared across source and target data. The open set domain adaptation shares some similarity with ZSL. This paper surveys the main-stream *closed-set* domain adaptation and transfer learning challenges.

3 Instance Re-weighting Adaptation

When the training and test data are drawn from different distribution, this is commonly referred to as *sample selection bias* or *covariate shift* [96]. Instance re-weighting aims to infer the resampling weight directly by feature distribution matching across different domains in a non-parametric manner. Generally, given a dataset $(x, y) \sim P_r(x, y)$, a learning model can be obtained by minimizing the following expected risk of the training set,

$$R[P_r, \theta, l(x, y, \theta)] = E_{(x,y)\sim P_r(x,y)}[l(x, y, \theta)] \tag{4}$$

But actually, we are more concerned about the expected risk of the testing set, shown as follows

$$\begin{aligned} R[P'_r, \theta, l(x, y, \theta)] &= E_{(x,y)\sim P'_r(x,y)}[l(x, y, \theta)] \\ &= E_{(x,y)\sim P_r(x,y)}[\frac{P'_r(x,y)}{P_r(x,y)} l(x, y, \theta)] \\ &= E_{(x,y)\sim P_r(x,y)}[\beta(x, y) l(x, y, \theta)] \end{aligned} \tag{5}$$

where $P_r(x, y)$ and $P'_r(x, y)$ represent the probability distribution of training and testing data, respectively. $l(x, y, \theta)$ is the loss function and $\beta(x, y)$ is the ratio between the two probabilities, which is amount to the weighting coefficient. Obviously, when $P_r(x, y) = P'_r(x, y)$, we have $\beta(x, y) = 1$.

From Eq.(5), we know that $P_r(x, y)$ and $P'_r(x, y)$ can be estimated for computing the weight $\beta(x, y)$ by following [97] based on the prior knowledge of the class distributions. Although this is intuitive, it requires very good density estimation of $P_r(x, y)$ and $P'_r(x, y)$. Particularly, a serious overweighting of the observations with very large coefficients $\beta(x, y)$ will be resulted from possible small errors or noise in estimating $P_r(x, y)$ and $P'_r(x, y)$. Therefore, in order to improve the reliability of the weights, $\beta(x, y)$ can be directly estimated by imposing flexible constraints into the learning model without having to estimate the two probability distributions.

Sample re-weighting based domain adaptation methods mainly focuses on the case where the difference between the source domain and the target domain is not too large. The objective is to re-weight the source samples so that the source data distribution can be more close to the target data distribution. Usually, when the distribution difference between the two domains is relatively large, the sample re-weighting methods can be combined with others (e.g. feature adaptation) for auxiliary transfer learning. Instance re-weighting has been studied with different models, which can be divided into three categories based on weighting scheme: (i) *Intuitive Weighting*, (ii) *Kernel Mapping Based Weighting*, and (iii) *Co-training Based Weighting*. This kind of methods put emphasis on the learning or computation of the weights by using different criterions and training protocols. The taxonomy of instance re-weighting based models is summarized in Table 1.

TABLE 1
Our Taxonomy of Instance Re-weighting Adaptation Approaches

Re-weighting Adaptation	Model Basis	Reference
Intuitive Weighting	Adaptive tuning	[98], [99], [100], [101]
Kernel Map-Based		
Distribution Matching	KMM&MMD	[96], [102], [103]
Sample Selection	K-Means &l_{21}-norm	[104], [105]
Co-training-Based	Double classifiers	[106], [107]

3.1 Intuitive Weighting

Instance re-weighting based domain adaptation was first proposed for natural language processing (NLP) [98], [99]. In [98], Jiang and Zhai proposed an intuitive instance weighted domain adaptation framework, which introduced four parameters for characterizing the distribution difference between source and target samples. For example, for each $(x_i^s, y_i^s) \in \mathcal{D}_s$, the labeled source data, the parameter α_i that was used to indicate how likely $\mathcal{P}_{target}(y_i^s|x_i^s)$ is close to $\mathcal{P}_{source}(y_i^s|x_i^s)$ and the parameter β_i that was ideally computed as $\frac{\mathcal{P}_{target}(x_i^s)}{\mathcal{P}_{source}(x_i^s)}$ were introduced. Obviously, large α_i means the high confidence of the labeled source sample (x_i^s, y_i^s) contributing positively to the learning effectiveness. Small α_i means the two probabilities are very different, and the instance (x_i^s, y_i^s) can be *discarded* in the learning process. Additionally, for each $x_i^{t,u} \in \mathcal{D}_{t,u}$, the unlabeled target data, and for each possible label $y \in \mathcal{Y}$, the hypothesis space, the parameter $\gamma_i(y)$ that indicates how likely a tentative pseudo-label y can be assigned to $x_i^{t,u}$, then the $(x_i^{t,u}, y)$ is *included* as a training sample.

Generally, α_i and γ_i play an intuitive role in sample selection by removing those misleading source samples and adding those valuable labeled target samples during the transfer learning process. Although the optimal weighting values of these parameters for the target domain are unknown, the intuitions behind the weights can be served as guidelines for researchers designing heuristic parameter tuning scheme [98]. Therefore, adaptive learning of these intuitive weights remains still a challenging issue.

In [99], Wang et al. proposed two instance weighting schemes for neural machine translation (NMT) domain adaptation, i.e., sentence weighting and dynamic domain weighting. Specifically, given the parallel training corpus $\mathcal{D} = [\mathcal{D}_{in}, \mathcal{D}_o]$ consisting of in-domain data and out-of-domain data, the sentence weighted NMT objective function was written as

$$\mathcal{J}_{sw} = \sum_{\langle x_i, y_i \rangle \in \mathcal{D}} \lambda_i \log \mathcal{P}(y_i | x_i) \tag{6}$$

where λ_i is the weight to score each $\langle x_i, y_i \rangle$. $\mathcal{P}(\cdot)$ is the conditional probability activated by softmax function. x and y represent the source sentence and target sentence, respectively. For domain weighting (dw), a weight λ was designed for the in-domain data, and the NMT objective function Eq.(6) can be transformed as [99]

$$\mathcal{J}_{dw} = \lambda \sum_{\langle x, y \rangle \in \mathcal{D}_{in}} \log \mathcal{P}(y|x) + \sum_{\langle x', y' \rangle \in \mathcal{D}_o} \log \mathcal{P}(y'|x') \tag{7}$$

A dynamic batch weight tuning scheme was proposed by monotonically increasing the ratio of in-domain data in the minibatch, which is supervised by the training cost. Dai et al. proposed a TrAdaBoost [100] transfer learning method, which leveraged Boosting algorithm to automatically tune the weights of the training samples.

In [101], Chen et al. proposed a more intuitive weighting based subspace alignment method by re-weighting the source samples for generating source subspace that are close to the target subspace. Let $w = [w_1, \cdots, w_m]^T \in \mathcal{R}^m$ denote the weighting vector of the source samples. Obviously, the w_i w.r.t. the source sample x_i increases if its distribution is more close to target data. Therefore, a simple weight assignment strategy was presented for assigning larger weights to the source samples that are closer to target domain [101].

After obtaining the weight vector w, the weighted source space can be obtained by performing PCA on the following covariance matrix $\mathcal{C}$ of weighted source data,

$$\mathcal{C} = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)^T w_i (x_i - \mu) \tag{8}$$

where μ is the weighted mean vector. Then the eigenvectors P_S can span the source subspace. By performing PCA on the target data, the eigenvectors P_T can span the target subspace. Thereafter, the following unsupervised domain adaptation model, *subspace alignment* (SA) [108], with Frobenius norm minimization, was implemented.

$$\min_M \|P_S M - P_T\|_F^2 \tag{9}$$

The subspace alignment matrix M can be easily solved with least-square solution.

3.2 Kernel Mapping Based Weighting

The intuitive weighting based domain adaptation was implemented in the raw data space. In order to infer the sampling weights by distribution matching across source and target data in feature space in a non-parametric way, kernel mapping based weighting was proposed. Briefly, the distribution difference between source and target data can be better characterised by re-weighting the source samples such that the means of the source and target instances in a reproducing kernel Hilbert space (RKHS) are close [96]. Kernel mapping based weighting consists of two categories of methods: *Distribution Matching* [96], [102], [103] and *Sample Selection* [104], [105].

(1) Distribution Matching. The intuitive idea of distribution matching is to match the means between the source and target data in a reproducing kernel Hilbert space (RKHS) by resampling the weights of the source data. Two similar distribution matching criterions, i.e., kernel mean matching (KMM) [96] and maximum mean discrepancy (MMD) [109], [110], have been used as non-parametric statistic to measure the distribution difference. Specifically, Huang et al. [96] firstly proposed to re-weight the source samples with β, such that the KMM between the means of target data and the weighted source data is minimized.

$$\begin{aligned} &\min_{\beta} \|E_{x' \sim P'_r}[\Phi(x')] - E_{x \sim P_r}[\beta(x)\Phi(x)]\| \\ &s.t. \quad \beta(x) \geq 0, E_{x \sim P_r}[\beta(x)] = 1 \end{aligned} \tag{10}$$

where $\Phi(\cdot)$ is the nonlinear mapping function into RKHS.

Chu et al. [102] further proposed a selective transfer machine (STM) by minimizing KMM for distribution matching, and simultaneously minimizing the empirical risk of the classifiers learned on the reweighted training samples.

$$(w, s) = \arg\min_{w,s} R_w(\mathcal{D}^{tr}, s) + \lambda \Omega_s(\mathcal{D}^{tr}, \mathcal{D}^{te}) \tag{11}$$

where $R_w(\cdot)$ is the empirical risk (loss) on the training set $\mathcal{D}^{tr}$, Ω_s indicates the distribution mismatch formulated by KMM, s is the weighting vector of the source samples, and w is the classifier parameters. From Eq.(11), the KMM based distribution mismatch plays an important role in model regularization on the sampling weights.

More recently, Yan et al. [103] proposed a weighted MMD (WMMD) for domain adaptation, which was implemented with convolutional neural network. WMMD overcomes the flaw of conventional MMD that ignores the class weight bias and assumes the same class weights between source and target domain. WMMD is formulated as [103]

$$d_{wmmd}^2 = \| \frac{1}{\Sigma_{i=1}^M \alpha_{y_i^s}} \sum_{i=1}^{M} \alpha_{y_i^s} \phi(x_i^s) - \frac{1}{N} \sum_{j=1}^{N} \phi(x_j^t) \|_{\mathcal{H}}^2 \tag{12}$$

where $\alpha_{y_i^s}$ is the class weight w.r.t. the class y_i^s of the i^{th} source sample and $\phi(\cdot)$ is the nonlinear mapping into RHKS $\mathcal{H}$. M and N denote the number of samples drawn from source and target domain, respectively.

(2) Sample Selection is another kind of kernel mapping based re-weighting method. Zhong et al. [104] proposed a cluster based sample selection method KMapWeighted which was established on the assumption that the kernel mapping can make the marginal distribution across domains similar, but the conditional probabilities between two domains after kernel mapping are still different. Therefore, in the RKHS space, they further select those source samples that are more likely similar to target data via a K-means based clustering criterion. The data in the same cluster should be with the same labels and then the source samples with similar labels to target data were selected.

Long et al. [105] proposed a TJM method for domain adaptation method by minimizing the MMD based distribution mismatch between source and target data, in which the transformation matrix A was imposed with structural sparsity (i.e., $l_{2,1}$-norm regularization constraint) for sampling. Then, larger coefficients correspond to the strong correlation between the source samples and the target domain samples. The TJM model is provided as [105]

$$\min_{A^TMA=I} tr(A^TMA) + \lambda(\|A_s\|_{2,1} + \|A_t\|_F^2) \tag{13}$$

where the $l_{2,1}$-norm on source transformation A_s means that source outliers can be excluded in transferring to target domain, the target transformation A_t was regularized for smoothness, and $M = KHK^T$ is the deduced matrix from MMD. H is the centering matrix and K is kernel matrix.

3.3 Co-training Based Weighting

Co-training [66] assumes that the dataset is characterized into two different views, in which two classifiers are then separately learned for each view. The inputs with high confidence of one of the two classifiers can be moved to the training set. In weighting based transfer learning, Chen et al. proposed a CODA [106] method, in which two classifiers with different weight vectors were trained. For better training both classifiers on the training set, the two classifiers were jointly minimized with weighting. In essence, the method of sample re-weighting based on the classifier is similar to the TrAdaBoost [100] and KMapWeighted [104].

In [107], Chen et al. proposed a re-weighted adversarial adaptation network (RAAN) for unsupervised domain adaptation. Two classifiers including a multi-class source instance classifier $\mathcal{C}$ and a binary domain classifier $\mathcal{D}$ were designed for adversarial training. The domain classifier $\mathcal{D}$ aims to discriminate whether features are from source or target domain, while the domain feature representation network $\mathcal{T}$ tries to confuse them, which formulates an adversarial training manner. For improving the domain confusion effect, the source feature distribution is re-weighted with β during training of the domain classifier $\mathcal{D}$. With the gaming between $\mathcal{T}$ and $\mathcal{D}$ as GAN does [33], the following minimax objective function was used [107],

$$\min_{\mathcal{T}} \max_{\mathcal{D},\beta} \mathcal{L}_{adv}^{Re} \tag{14}$$

where the weight β is multiplied with $\mathcal{D}$, and both β and $\mathcal{D}$ were trained in a cooperative way. The learning of the source classifier $\mathcal{C}$ was easily performed by minimizing the cross-entropy loss.

3.4 Discussion and Summary

In this section, we recognize three kinds of instance re-weighting: intuitive, kernel mapping and co-training. The intuitive re-weighting advocates to tune the weights of the source samples, such that the weighted source distribution is closer to target distribution. The kernel mapping based re-weighting is further divided into distribution matching and sample selection. The former aims to learn source sample weights such that the kernel mean discrepancy between target data and the weighted source data is minimized, and the latter advocates sample selection by using K-means clustering (cluster assumption) and $l_{2,1}$-norm based structural sparsity in RKHS space. The co-training mechanism focus on learning with two classifiers. Additionally, the adversarial training of the weighted domain classifier can facilitate domain confusion.

Although instance re-weighting is the earliest method to address domain mismatch problem, there are still some directions worth studying: 1) essentially, instance weighting can be incorporated into most of learning frameworks; 2) the initialization and estimation of instance weights are important and can be treated as a latent variable obeying some probability distribution.

4 Feature Adaptation

Feature adaptation aims to discover the common feature representation of the data drawn from multiple sources by using different techniques including linear and nonlinear ones. In the past decade, feature adaptation induced transfer adaptation learning has been intensively studied, which, in our taxonomy, can be categorized into (i) *Feature Subspace*-Based, (ii) *Feature Transformation*-Based, (iii) *Feature Reconstruction*-Based and (iv) *Feature Coding*-Based. Despite these advances, the technical challenges being faced by researchers lie in the domain subspace alignment, projection learning for distribution matching, generic representation and shared domain dictionary coding. The taxonomy of feature adaptation challenges is summarized in Table 2.

TABLE 2
Our Taxonomy of Feature Adaptation Challenges

Feature Adaptation	Model Basis	Reference
Feature Subspace		
Geodesic path	Grassman manifold	[111], [112]
Alignment	Subspace learning	[108], [113], [114]
Feature Transformation		
Projection	MMD&HSIC& Bregman divergence	[115], [116], [117], [118]
Metric	First/second-order statistic	[119], [120], [121], [122]
Augmentation	Zero-padding& Generative	[123], [124], [125], [126]
Feature Reconstruction		
Low-rank models	Low-rank representation (LRR)	[127], [128], [126], [129]
Sparse models	Sparse subspace clustering (SSC)	[130], [131], [132], [133]
Feature Coding		
Domain-shared dictionary	Dictionary learning	[134], [135], [136], [137]
Domain-specific dictionary	Dictionary learning	[138], [139], [140], [141]

4.1 Feature Subspace-Based

Learning subspace generally resorts to unsupervised domain adaptation. Three representational models are referred to as sampling geodesic flow (SGF) [111], geodesic flow kernel (GFK) [112] and subspace alignment (SA) [108]. There exists a common property of the three methods, i.e. the data is assumed to be represented by a low-dimensional linear subspace. That is, a low-dimensional Grassmann manifold is embedded in the high-dimensional data. Generally, principal component analysis (PCA) was used to construct the Grassmann manifold, where the source and target domains become two points and a geodesic flow or path was formulated. SGF proposed by Gopalan, et al. [111] is an unsupervised low-dimensional subspace transfer method, which samples a group of subspaces along the geodesic path between source and target data, and aims to find an intermediate representation with closer domain distance.

Similar to but different from SGF, Gong et al. proposed a GFK [112], in which the geodesic flow kernel was used to model the domain shift by integrating an infinite number of subspaces. GFK explores an intrinsic low-dimensional spatial structure that associates two domains and the main idea behind is to find a geodesic line from $\phi(0)$ to $\phi(1)$, such that the raw feature can be transformed into a space of infinite dimension from $\phi(0)$ to $\phi(1)$ where distribution difference is easy to be reduced. In particular, the infinite dimensional features in the manifold space can be represented as $z = \phi(t)^T x$. The inner product of the transformed features z_i and z_j defines a positive semi-definite geodesic flow kernel as follows:

$$\left\langle z_i^{\infty}, z_i^{\infty} \right\rangle = \int_0^1 (\phi(t)^T x_i)^T (\phi(t)^T x_i) dt = x_i^T G x_j \qquad (15)$$

where G is a positive semi-definite mapping matrix. With $z = \sqrt{G}$x, features in the original space can be transformed into the Grassmann manifold space.

For aligning the source subspace to the target subspace, in SA [108], Fernando, et al. proposed to move closer the two subspaces with respect to the points in Grassmann manifold by directly designing an alignment matrix M, which well bridges the source and target subspaces. The model of SA is described in Eq.(9). As presented in SA, the subspaces of source and target data were spanned by the eigenvectors induced with a PCA. Further, Sun and Saenko proposed a subspace distribution alignment (SDA) [113] by simultaneously aligning the distributions as well as the subspace bases, which overcomes the flaw of SA that does not take into account the distribution difference.

More intuitively, Liu and Zhang proposed a guided transfer hashing (GTH) [114] framework, which introduced a more generic method for moving the source subspace W_s closer to target subspace W_t,

$$\min_{W_s, W_t} \frac{1}{2} \| M^{\frac{1}{2}} \odot (W_t - W_s) \|^2 \qquad (16)$$

where M is a weighting matrix on the difference between source and target subspaces. Through this way, the two subspaces can be solved alternatively and progressively, which is therefore recognized as a guided transfer mechanism.

4.2 Feature Transformation-Based

This kind of models aim to learn a transformation or projection of the data with some distribution matching metrics between source and target domains [5], [142], [143]. Then, the transformed or projected feature distribution difference across two domains can be removed or relieved. Feature transformation based domain adaptation has been a mainstream in visual transfer learning community in last years, which can be further divided into *Projection*, *Metric*, and *Augmentation* according to the model formulation.

(1) Projection-Based domain adaptation aims to solve a projection matrix in source and target domain for reducing the marginal distribution difference and conditional distribution difference between domains, by introducing *Kernel Matching Criterion* [118], [115], [144], [116], [117], [145], [146], [147] and *Discriminative Criterion* [148], [149], [150], [151]. The kernel matching criterion generally adopts the maximum mean discrepancy (MMD) statistic, which characterizes the *marginal* distribution difference and *conditional* distribution difference between source and target data. In unsupervised domain adaptation setting, the labels of target domain samples are generally unavailable, therefore the pseudo-labels of target samples should be iteratively predicted for quantifying the conditional MMD between domains [148], [152]. The discriminative criterion focus on within-class compactness and between-class separability of the projection. Mathematically, the formulation of empirical nonparametric MMD in universal RKHS is written as

$$d_{\mathcal{H}}^2(\mathcal{D}_s, \mathcal{D}_t) = \| \frac{1}{M} \sum_{i=1}^{M} \phi(x_i^s) - \frac{1}{N} \sum_{j=1}^{N} \phi(x_j^t) \|_{\mathcal{H}}^2 \qquad (17)$$

Specifically, with MMD based kernel matching criterion, Pan and Yang firstly proposed a transfer component analysis (TCA) [115] by introducing the marginal MMD with projection as the loss function. The joint distribution adaptation (JDA) proposed by Long et al. [116] further introduced the conditional MMD on the basis of TCA, such that the cross-domain distribution alignment becomes more discriminative. The general model can be written as

$$\min_W d_m^2(X_S, X_T, W) + \lambda d_c^2(X_S, X_T, Y_S, Y_T', W) \qquad (18)$$

where W denotes the projection matrix, Y_T' denotes the predicted pseudo-label of target data, d_m^2 and d_c^2 represent the marginal and conditional distribution discrepancy, respectively. For improving the discrimination of the projection matrix, such that the within-class compactness and between-class separability in each domain can be better characterized, the model with joint discriminative subspace learning and MMD minimization was proposed, for example, JGSA [148] and CDSL [149], and generally written as

$$\min_W F(W, X_S, X_T, Y_S, Y_T') + \lambda d_{\{m,c\}}^2(X_S, X_T, W) \qquad (19)$$

where $F(\cdot)$ is a scalable subspace learning function of the projection W, for example, linear discriminative analysis (LDA), local preservation projection (LPP), marginal fisher analysis (MFA), principal component analysis (PCA), etc. In addition to the MMD based criterion in projection based transfer model, Bregman divergence based [118], Hilbert-Schmidt independence criterion (HSIC) based [153], [117], [154], [133], and manifold criterion based [126].

In [118], Si et al. proposed a transfer subspace learning (TSL) by introducing a Bregman divergence-based discrepancy as regularization instead of MMD, which is written as

$$W = arg \min_W F(W) + \lambda D_W(P_L \| P_U) \qquad (20)$$

where $F(W)$ is similar to Eq.(19) and $D_W(P_L||P_U)$ is the Bregman divergence-based regularization that measures the distance between the probability distribution of training samples P_L and that of the testing samples P_U in the projected subspace W.

The HSIC proposed by Gretton et al. [153], the same author as that of MMD, was used to measure the dependency between two sets $\mathcal{X}$ and $\mathcal{Y}$. Let k_x and k_y denote the kernel function w.r.t. the RKHS $\mathcal{F}$ and $\mathcal{G}$. The HSIC is mathematically written as [153]

$$\begin{aligned} & HSIC(\mathcal{X},\mathcal{Y},\mathcal{F},\mathcal{G}) \\ & = \|C_{\mathcal{X}\mathcal{Y}}\|^2_{H-S} = (N-1)^{-2}Tr(\mathcal{K}_{\mathcal{X}}H\mathcal{K}_{\mathcal{Y}}H) \\ & s.t. \quad H = I - N^{-1}1_{N\times 1}1^T_{N\times 1} \end{aligned} \tag{21}$$

where N is the size of the set $\mathcal{X}$ and $\mathcal{Y}$ and $\|C_{\mathcal{X}\mathcal{Y}}\|^2_{H-S}$ is the Hilbert-Schmidt norm of the cross-covariance operator. $\mathcal{K}_{\mathcal{X}}$ and $\mathcal{K}_{\mathcal{Y}}$ denote the two kernel Gram matrix, and H is the centering matrix. HSIC will be zero if and only if $\mathcal{X}$ and $\mathcal{Y}$ are independent. In [133], Wang et al. proposed to use the projected HSIC as regularization, which is written as

$$\min_W F(W) - \lambda HSIC(W,\mathcal{X},\mathcal{Y},\mathcal{F},\mathcal{G}) \tag{22}$$

where $\mathcal{Y}$ denotes the label set of source and target data. Obviously, the model constrains W to reduce the independency between feature set $\mathcal{X}$ and label set $\mathcal{Y}$, such that the classification performance can be improved. In model formulation, the general way is to set a common projection for both domains. Another way is to learn two projections W_S and W_T, one for each domain, such that domain specific projection can be solved [148], [122], [114], [129]. For moving the two projections of both domains closer, the Frobenius norm of their difference like Eq.(16) can be used.

(2) Metric-Based aims to learn a good distance metric from labeled source data which can be easily adapted to a related but different target domain [155]. Metric transfer has a close link to projection based examples, if the metric M is a semi-definite matrix and can be decomposed into $M = WW^T$ [119]. The metric-based transfer can be divided into *First-order statistic* [119], [120], [142], [156], [157], [158] and *Second-order statistic* [121], [122], [159], [160], [161] based distance metric, such as Euclidean or Mahalanobis distance.

The First-order metric transfer generally learns a metric under which the distance between source and target feature is minimized, and it can be written as

$$\min_M d(M, \phi(X_S), \phi(X_T)) + \lambda\Re(M) \tag{23}$$

where $\phi(\cdot)$ is the feature representation or mapping function, and it can be linear mapping [142], kernel mapping [120], [157], auto-encoder [119] or neural network [156].

For example, the robust transfer metric learning (RTML) proposed by Ding et al. [119] adopted an auto-encoder based feature representation for metric learning, such that the Mahalanobis distance between source and target domain is minimized. The objective function of RTML is as follows:

$$\min_{M\in S^d_+} \sum_{i=0}^{c} tr(\phi_i M) + \alpha\left\|\overline{X} - M\widetilde{X}\right\|_F^2 + \lambda rank(M) \tag{24}$$

where M is positive semi-definite matrix, $\overline{X}$ is the repeated version of X, $\widetilde{X}$ is the randomly corrupted version of $\overline{X}$. The first item is Mahalanobis distance induced domain discrepancy under metric M, the second item is auto-encoder for feature learning, and the third term is the low-rank constraint for characterizing the internal correlation between domains.

The Second-order metric transfer generally learns a metric under which the distance between the covariances of source and target domain instead of the means is minimized [160], [122], [159], [121]. For example, Sun et al. [159], [121] proposed a simple but efficient correlation alignment (CORAL) by aligning the second-order statistic (i.e. the covariance) between source and target distributions instead of the first-order metric. By introducing a metric matrix A, the difference between source covariance Σ_S and target covariance Σ_T in CORAL can be minimized by solving

$$\min_A \|A^T\Sigma_S A - \Sigma_T\|_F^2 \tag{25}$$

The Eq.(25) is amount to matching the two centered Gaussian distribution, which is the basic assumption for such second-order statistic based transfer.

(3) Augmentation-Based domain adaptation often assume that the feature representation is grouped with three types: common representation, source-specific representation and target-specific representation. In general case, the source domain should be characterized as the composition of common component and source-specific component, and similarly, the target domain should be characterized as the composition of common component and target-specific component. Feature augmentation based DA can be divided into the generic *Zero Padding* [162], [123], [163], [124], [164] and the latest *Generative* [125], [126] types.

Zero Padding was firstly proposed by Daume III [162], which presented an EasyAdapt (EA) model. Assume the raw input data space to be $\mathcal{X} \in \Re^F$, then the augmented feature spaces should be $\mathcal{Y} \in \Re^{3F}$. By defining the mapping functions of source and target domain from $\mathcal{X}$ to $\mathcal{Y}$ as $\Phi_s(\cdot)$ and $\Phi_t(\cdot)$, respectively. Then, there is

$$\Phi_s(x) = [x, x, 0], \Phi_t(x) = [x, 0, x] \tag{26}$$

where $0 \in \Re^F$ is a zero vector. The first, second and third bits of the augmented feature $\Phi(x)$ in Eq.(26) represent the common, source-specific and target-specific feature component, respectively. However, in heterogeneous domain adaptation that addressing different feature dimensions between source and target domain [165], [166], [167], for example, cross-modal learning (e.g., images vs. text), Li et al. [124] argued that such simple zero-padding for dimensionality consistence between domains is not meaningful. The reason is that there would be no correspondences between the heterogeneous features. Therefore, Li et al. [124] proposed a heterogeneous feature augmentation (HFA) model, which incorporates the projected features together with the raw features for feature augmentation by introducing two projection matrices $P \in \Re^{d_c\times d_s}$ and $Q \in \Re^{d_c\times d_t}$. The augmented feature for source and target domain can be written as

$$\Phi_s(x_s) = [Px_s, x_s, 0_{d_s}], \Phi_t(x_t) = [Qx_t, 0_{d_t}, x_t] \tag{27}$$

where d_s and d_t represent the dimensionality of source and target data, respectively. For incorporating the unlabeled target data, Daume III further proposed an EA++ model with zero padding based feature augmentation for semi-supervised domain adaptation [123], [163]. Chen et al. [164] proposed to use zero padding based camera correlation aware feature augmentation (CRAFT) for cross-view person re-identification.

Generative methods for feature augmentation mainly focus on plausible data generation for enhancing the robustness for domain transfer. In [125], Volpi et al. proposed an adversarial feature augmentation by introducing two generative adversarial nets (GANs). The first GAN was used to train the generator S for synthesizing implausible source images (data augmentation) by inputting noise and conditional labels. The second GAN was used to train the shared feature encoder E (feature augmentation) for both domains, by adversarial learning with the synthesized source images via S. Finally, the encoder E was used as the domain adapted feature extractor shared by both domains. In [126], Zhang et al. proposed a manifold criterion guided intermediate domain generation for feature augmentation, which improved the transfer performance by generating high-quality intermediate features.

4.3 Feature Reconstruction-Based

Feature reconstruction between source and target data using a representational matrix for domain transfer has been studied for several years. By linear sample reconstruction in an intermediate representation with low-rankness and sparsity, it can well characterize the intrinsic relatedness and correspondences between source and target domain, while excluding noises and outliers during domain adaptation. To this end, feature reconstruction based domain transfer can be generally divided into two types: *Low-rank Reconstruction* [127], [128], [126], [129] and *Sparse Reconstruction* [130], [132], [131], [133]. For the former, for characterizing the domain differences and uncovering the domain noises, the reconstruction matrix was imposed with low-rank constraint, such that the relatedness between domains can be discovered. For the latter, sparsity or structural sparsity was generally used for transferrable sample selection. Methodologically, reconstruction based domain transfer is closely related to low-rank representation (LRR) [168], [169], matrix recovery [170], [171] and sparse subspace clustering (SSC) [172], [173], [174].

(1) Low-rank Reconstruction based domain adaptation was firstly proposed by Jhuo et al. [127], in which the W transformed source feature was reconstructed by the target domain with low-rank constraint on the reconstruction matrix and $l_{2,1}$-norm constraint on the error.

$$\begin{aligned} &\min_{W,Z,E} rank(Z) + \alpha\|E\|_{2,1} \\ &s.t. \quad WX_S = X_TZ + E, WW^T = I \end{aligned} \tag{28}$$

However, seeking for an alignment between WX_S and X_T may not transfer knowledge directly, due to the out of domain problem of W for unilateral projection.

On the basis of [127], Shao et al. [128] proposed a latent subspace transfer learning (LTSL), which tends to reconstruct the target data by using the source data as basis in a projected latent subspace.

$$\begin{aligned} &\min_{W,Z,E} F(W, X_S) + \lambda_1 rank(Z) + \alpha\|E\|_{2,1} \\ &s.t. \quad W^TX_S = W^TX_TZ + E \end{aligned} \tag{29}$$

where $F(\cdot)$ is a subspace learning function, similar to Eq.(19), Eq.(20) and Eq.(22). By comparing Eq.(28) to Eq.(29), the major difference lies in the latent space learning of W for both domains in LTSL. Both methods, established on LRR, advocated low-rank reconstruction between domains for transfer learning. As demonstrated in [169], trivial solution may be easily encountered when handling disjoint subspaces and insufficient data using LRR and a strong independent subspace assumption is necessary.

(2) Sparse Reconstruction based domain transfer was established on the SSC, which, different from LRR, is well supported by theoretical analysis and experiments when handling the data near the intersections of subspaces [173]. Therefore, in [130], Zhang et al. proposed a latent sparse domain transfer (LSDT) model, which jointly learn the sparse coding Z between domains and the latent subspace W.

$$\begin{aligned} \min_{Z,W} \quad & \|Z\|_1 + \lambda_1\|WX_T - WXZ\|_F^2 \\ & + \lambda_2\|X - W^TWX\|_F^2 \\ s.t. \quad & WW^T = I, \quad \mathbf{1}^T_{N_S+N_T}Z = \mathbf{1}^T_{N_T}, \quad Z_{N_S+i,i} = 0, \\ & \forall i = 1, ..., N_T \end{aligned} \tag{30}$$

where X is the feature set grouped by X_S and X_T.

With the sparsity constraint on Z, the most transferrable samples can be selected during domain adaptation, which is more robust to noise or outliers drawn from source domain. The model has also been kernerlized by defining the projection W as the linear representation of X. The reconstruction is then implemented in a high-dimensional reproducing kernel Hilbert space (RKHS), based on the Representor theorem. In [132], Zhang et al. proposed a $l_{2,1}$-norm constraint based reconstruction transfer model with discriminative subspace learning and the domain-class consistency was guaranteed. The joint constraint with low-rankness and sparsity for the reconstruction matrix was proposed in [131], such that the global and local structures of data can be preserved.

4.4 Feature Coding-Based

In feature reconstruction based transfer models, the focus is the learning of reconstruction coefficients across domains, on the basis of the raw feature of source or target data. Different from that, feature coding based transfer learning put emphasis on seeking a group of basis (i.e., dictionary) and representation coefficients in each domain, which was generally called domain adaptive dictionary learning. The typical dictionary learning approach aims to minimize the representation error of the given data set under a sparsity constraint [175], [176], [177]. The cross-domain dictionary learning aims to learn domain adaptive dictionaries without requiring any explicit correspondences between domains, which was generally divided into two types of learning, *domain-shared dictionary*-based [134], [135], [136], [137] and *domain-specific dictionary*-based [138], [139], [140], [141],

[178]. Obviously, the former resorts to learning one common dictionary for both domains, while the latter contributes to obtain two or more dictionaries for each domain.

(1) Domain-shared dictionary aims at representing the source and target domain using a common dictionary. In [134], [136], Shekhar et al. proposed to separately represent the source and target data in a latent subspace with a shared dictionary D, which can be written as

$$\min_{D,P,\alpha} \sum_{k\in\{s,t\}} \|P_{(k)}X_{(k)} - D\alpha_{(k)}\|_F^2 + \Re(D,P,\alpha) \tag{31}$$

where P denotes the latent subspace projection, α denotes the representational coefficients for source data X_s and target data X_t using a shared dictionary D, and $\Re(\cdot)$ denotes the regularizer. The shared dictionary D is demonstrated to incorporate the common information from both domains.

(2) Domain-specific dictionary tends to learn multiple dictionaries, one for each domain, to represent the data in each domain based on domain specific or common representation coefficients [140], [178]. The general model can be written as

$$\min_{D,P,\alpha} \sum_{k\in\{s,t\}} \|X_{(k)} - D_{(k)}\alpha_{(k)}\|_F^2 + \Omega(\alpha_s,\alpha_t) \tag{32}$$

where $\Omega(\cdot)$ denotes the difference between representation coefficients of source and target. If $\alpha_s = \alpha_t = \alpha$, then $\Omega(\alpha_s,\alpha_t) = 0$ and the model in Eq.(32) is degenerated as the common representation coefficients based domain adaptive dictionary learning [138].

In [139], [135], [141], a set of intermediate domains that bridge the gap between source and target domains were incorporated as multiple dictionaries $\{D_k\}_{k=1}^{K-1}$, which can progressively capture the intrinsic domain shift between source domain dictionary D_0 and target domain dictionary D_K. The difference $\triangle D_k$ between the atoms of adjacent two sub-dictionaries can well characterize the incremental transition and shift between two domains. Actually, this kind of models can be linked with SGF [111] and GFK [112] by sampling finite or infinite number of intermediate subspaces on the Grassmann manifold for better capturing the intrinsic domain shift.

4.5 Discussion and Summary

In this section, feature adaptation methods are presented, including subspace, transformation, reconstruction and coding based types. Feature subspace focuses on the subspace alignment between domains in Grassmann manifold. Feature transformation is further categorized into three subclasses: projection learning with MMD criterion, metric learning with first-order or second-order statistics and augmentation with zero-padding. Feature reconstruction aims to explicitly bridge the source and target data in a latent subspace by low-rank or sparse reconstruction. Finally, the feature coding focus on domain data representation by learning domain adaptive dictionaries without explicit correspondences between domains.

Feature adaptation is intensively studied and two future directions are specified: 1) more reliable probability distribution similarity metric is needed, except the Gaussian kernel induced MMD; 2) for learning domain-invariant representation, model ensemble of linear and nonlinear ones is desired.

5 Classifier Adaptation

In cross-domain visual categorization, classifier adaptation based TAL aims to learn a generic classifier by leveraging labeled samples drawn from source domain and few labeled samples from target domain [3], [179], [180], [181]. Typical cross-domain classifier adaptation can be divided into (i) *Kernel Classifier*-Based [3], [182], [183], [184], [179], [185], [186], (ii) *Manifold Regularizer*-Based [187], [188], [189], [190], [191] and (iii) *Bayesian Classifier*-Based [192], [193], [194], [195], [196], [197]. The taxonomy of classifier adaptation approaches is summarized in Table 3.

TABLE 3
Our Taxonomy of Classifier Adaptation Challenges

Classifier Adaptation	Model Basis	Reference
Kernel Classifier	SVM&MKL	[3], [183], [184], [179], [185], [186]
Manifold Regularizer	Label Propagation &MMD	[187], [188], [189], [190], [191]
Bayesian Classifier	Probabilistic graph models	[192], [193], [194], [195], [196], [197]

5.1 Kernel Classifier-Based

Yang et al. [3] firstly proposed an adaptive support vector machine (ASVM) in 2007 for target classifier training, which assumed that there exists a bias $\Delta f(x)$ between source classifier $f^a(x)$ and target classifier $f(x)$. This means that the bias can be added to the source classifier to generate a new decision function, that is adapted to classifying the target data. There is,

$$f(x) = f^a(x) + \Delta f(x) = f^a(x) + w^T\phi(x) \tag{33}$$

where w is the parameter of the bias function $\Delta f(x)$, which was solved by standard SVM,

$$\begin{aligned} &\min_w \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\varepsilon_i \\ &s.t. \quad y_i f^a(x_i) + y_i w^T\phi(x_i) \geq 1-\varepsilon_i, \quad \varepsilon_i > 0 \end{aligned} \tag{34}$$

In Eq.(34), $f^a(\cdot)$ was known and trained on labeled source data, (x_i, y_i) are drawn from few labeled target data, and w is the parameter of $\Delta f(\cdot)$ rather than $f(\cdot)$.

More recently, on the basis of ASVM, Duan et al. proposed a series of multiple kernel learning (MKL) based domain transfer classifiers [182], [183], [184], [185], including AMKL, DTSVM, and DTMKL, in which the kernel function was assumed to be a linear combination of multiple predefined base kernel functions by following the MKL methodology [198], [199]. Additionally, for reducing the domain distribution mismatch, MMD based kernel matching metric $d_{\mathbb{k}}^2(\cdot)$ was jointly minimized with the structural risk based classifiers. The general model of MKL based classifier adaptation can be written as

$$\min_{\mathbb{k},f} R(\mathbb{k}, f, X_S, X_T) + \lambda\Omega(d_{\mathbb{k}}^2(X_S, X_T)) \tag{35}$$

where $R(\cdot)$ denotes the structural risk on labeled training samples, f is the decision function, $\Omega(\cdot)$ is the monotonic increasing function, $\Bbbk = \sum_{m=1}^{M} d_m k_m$ is a linear combination of a set of base kernels $k_m s$ with $\sum_{m=1}^{M} d_m = 1$ and $d_m \geqslant 0$. The structural risk $R(\cdot)$ was generally formulated based on the hinge loss, i.e., $l_\hbar(t) = \max(0, 1 - t)$, as that in SVM. Duan et al. [200] also proposed a domain adaptation machine (DAM), which incorporated SVM hinge loss based structural risk with multiple domain regularizers for target classifier learning. Regularized least-square loss based classifier adaptation can be referred to as [187], [188].

5.2 Manifold Regularizer-Based

The manifold assumption in semi-supervised learning means that the the similar samples with small distance in feature space more likely belongs to the same class. By constructing the affinity graph based manifold regularizer, under which, the classifier trained on source data can be more easily adapted to target data through label propagation. Long et al. [187] and Cao et al. [190] proposed ARTL and DMM which advocated manifold regularization based structural risk and between-domain MMD minimization for classifier training, structural preservation and domain alignment. In [191], Yao et al. proposed to simultaneously minimize the classification error, preserve the geometric structure of data and restrict similarity characterized on unlabeled target data. Zhang and Zhang [188] proposed a manifold regularization based least-square classifier EDA on both domains with label pre-computation and refining for domain adaptation. More recently, Wang et al. [189] proposed a domain-invariant classifier MEDA in Grassmann manifold with structural risk minimization, while performing cross-domain distribution alignment of marginal and conditional distributions with different importances. Graph based manifold regularization $\mathcal{M}(\cdot)$ can be written as

$$\mathcal{M}(X) = \sum_{i,j} W_{ij}(f(x_i) - f(x_j))^2 = tr(F^T \mathcal{L} F) \tag{36}$$

where X is the data of source and target domain, F is the predicted labels, $\mathcal{L} = D - W$ is the Laplacian matrix, W_{ij} is the weight between sample i and j, and D is a diagonal matrix with $D_{ii} = \sum_i W_{ij}$. This term constrains the geometric structure preservation in label propagation and helps classifier adaptation. Although manifold regularizer can improve classifier adaptation performance, the fact is that the manifold assumption may not always hold, particularly when domain distribution does not match [126].

5.3 Bayesian Classifier-Based

In learning complex systems with limited data, Bayesian learning can well integrate prior knowledge to improve the weak generalization of models caused by data scarcity. For unsupervised domain adaptation, an underlying assumption in the kernel classifier and manifold classifier based models is that the conditional domain shift between domains can be minimized without relying on the target labels. Additionally, these methods are deterministic, which rely more on the expensive cross-validation for determining the underlying manifold space where the kernel mismatch between domains is effectively reduced. Recently, probabilistic model, i.e., Bayesian classifier based graphical models for DA/TL have been studied [192], [193], [194], [195], [196], [197], which aim to have better insights on the transferrable process from source domain to target domain.

In [192], Gönen and Margolin firstly proposed graphical model, i.e., kernelized Bayesian transfer learning (KBTL) for domain adaptation. This work aims to seek a shared subspace and learn a coupled linear classifier in this subspace using a full Bayesian framework, solved by a variational approximation based inference algorithm. In [195], Gholami et al. proposed a probabilistic latent variable model (PUnDA) for unsupervised domain adaptation, by simultaneously learning the classifier in a projected latent space and minimizing the MMD based domain disparity. A regularized Variational Bayesian (VB) algorithm was used for efficient model parameter estimation in PUnDA, because the computation of exact posterior distribution of the latent variables is intractable. More recently, Karbalayghareh et al. [196] proposed an optimal Bayesian transfer learning (OBTL) classifier to formulate the optimal Bayesian classifier (OBC) in target domain by using the prior knowledge of source and target domains, where OBC [201] aims to achieve Bayesian minimum mean squared error over uncertainty classes. In order to avoid costly computations such as MCMC sampling, OBTL classifier was derived based on the Laplace approximated hypergeometric functions.

5.4 Discussion and Summary

In this section, classifier adaptation including kernel classifier, manifold regularizer and Bayesian classifier are surveyed, which mostly rely on a small amount of tagged target domain data and facilitate semi-supervised transfer learning. This can be easily adapted to unsupervised transfer learning by pre-computing and iteratively updating the pseudo-labels of the completely unlabeled target domain in classifier adaptation. The kernel classifier focuses on SVM or MKL learning jointly with MMD based domain disparity minimization. The manifold regularizer based models aim to preserve the data affinity structure for label propagation. The Bayesian classifier based models resort to compensating the generalization performance loss due to data scarcity by modeling on the prior knowledge under reliable distribution assumptions, and having theoretical understanding on transfer learning from the viewpoint of data generation.

However, some inherent flaws exist: 1) incorrect pseudo-labels of target data significantly lead to performance degradation; 2) inaccurate distribution assumption in estimating various latent variables produces very negative effect; 3) the manifold assumption between domains does not hold for serious domain disparity.

6 Deep Network Adaptation

Deep neural networks (DNNs) have been recognized as dominant techniques for addressing computer vision tasks, due to their powerful feature representation and end-to-end training capability. Although DNNs can achieve more generalized features and performance in visual categorization, they rely on massive amounts of labeled data. For a target domain where the labeled data is unavailable or a very few labeled data is available, deep network adaptation started to rise. Yosinski et al. [202] has discussed the transferability of features in bottom, middle and top layers of DNNs, and demonstrated that the transferability of features decreases as the distance between domains increases. In [203], Donahue et al. proposed the deep convolutional activation feature (DeCAF) extracted by using a pre-trained AlexNet model [16], which has well proved the generalization of DNNs for generic visual classification. This work further facilitated deep transfer learning and deep domain adaptation. Generally, the presented three types of TAL models in Section 3, 4 and 5, including instance re-weighting, feature adaptation and classifier adaptation, can be incorporated into DNNs with end-to-end training for deep network adaptation. In 2015, Long et al. [204], [205] proposed a deep adaptation network (DAN) for learning transferrable features, which, for the first time opened the topic of deep transfer and adaptation. The basic idea of DAN is to enhance feature transferability in task-specific layers of DNNs by embedding the higher layered features into reproducing kernel Hilbert spaces (RKHSs) for nonparametric kernel matching (e.g., MMD-based) between domains. In training process, DAN was trained by fine-tuning on the ImageNet pre-trained DNN, such as AlexNet [16], VGGNet [206], GoogLeNet [207] and ResNet [17]. Currently, the works in deep network adaptation can be divided into (i) *Marginal Alignment-Based*, (ii) *Conditional Alignment-Based* and (iii) *Autoencoder-Based*, in which the first two focus on convolutional neural networks. The taxonomy of deep network adaptation challenges is summarized in Table 4.

TABLE 4
Our Taxonomy of Deep Network Adaptation Challenges

DEEP NET ADAPTATION	MODEL BASIS	REFERENCE
Marginal Alignment	CNN&MMD	[204], [208], [209], [210], [211], [212]
Conditional Alignment	CNN&MMD &Semantics	[205], [213], [214]
Autoencoder-Based	Stacked Denoising autoencoders	[215], [216], [217], [166], [218], [219]

6.1 Marginal Alignment-Based

In unsupervised deep domain adaptation frameworks, for reducing the distribution disparity between labeled source domain and unlabeled target domain, the top layered features were generally transformed to a RKHS space where the maximum mean discrepancy (MMD) based kernel matching between domains was performed, which is recognized as marginal alignment based deep network adaptation [204], [208], [209], [210]. For image classification, the softmax guided cross-entropy loss on the labeled source data is generally minimized. Representative works can be referred to as DDC proposed by Tzeng et al. [208] and DAN [204]. The model can be written as

$$\min_{\Theta} \frac{1}{N_s}\sum_{i=1}^{N_s}\mathcal{J}(\theta(x_i), y_i) + \lambda\sum_{l} d_{ma}^2(\mathcal{D}_s^l, \mathcal{D}_t^l) \tag{37}$$

where $\mathcal{J}(\cdot)$ is the cross-entropy loss function, $\theta(\cdot)$ is the feature representation function, $\mathcal{D}^l$ denotes the domain feature set from the l^{th} layer and $d_{ma}^2(\cdot)$ is the marginal alignment function (i.e., MMD in Eq.(17)) between domains. Clearly, in Eq.(37), multiple MMDs were formulated, one for each layer, and the summation of all MMDs is minimized. For better measuring the discrepancy between domains, a unified MMD called joint MMD (JMMD) was further designed by Long et al. [210] in a tensor product Hilbert space for matching the joint distribution of activations of multiple layers.

The model in Eq.(37) does not take into account the network outputs of target domain stream, which may not well adapt the source classifier to target data. For addressing this problem, conditional-entropy minimization principle [220] that favors the low-density separation between classes in unlabeled target data $\mathcal{D}_t$ was further exploited in [205], [211], [212]. The entropy minimization is written as

$$\min_{f\in\mathcal{F}} -\frac{1}{N_t}\sum_{i=1}^{N_t}\sum_{j=1}^{C} f_j(x_i^t)\log f_j(x_i^t) \tag{38}$$

where $f_j(x)$ is the probability that sample x is predicted as class j. Entropy minimization is amount to *uncertainty* minimization of the predicted labels of target samples. Additionally, by following the assumption of ASVM in [3], the residual $\Delta f(x)$ between source and target classifiers was learned in the residual transfer network (RTN) [211], with a residual connection.

6.2 Conditional Alignment-Based

In marginal alignment based deep network adaptation, only the top layered feature matching in RKSH spaces was formulated by using the nonparametric MMD metric. However, the high-level semantic information was not taken into account in domain alignment, which may degrade the adaptability of source data trained DNNs to unlabeled target domain. Therefore, conditional alignment based deep network adaptation methods were presented jointly with marginal alignment based models [213], [214]. Similar to the formulation of MMD in Eq.(17), the conditional alignment was generally formulated by building MMD like metric d_{ca}^2 on the probabilities p, the uncertainty that predicts a sample to class c between domains.

$$d_{ca}^2 = \sum_{c=1}^{C}\|\frac{1}{n_s}\sum_{i=1}^{n_s}p(y_i^s = c|x_i^s) - \frac{1}{n_t}\sum_{j=1}^{n_t}p(y_i^t = c|x_j^t)\|^2 \tag{39}$$

Therefore, conditional alignment based deep adaptation model was generally constructed by combining Eq.(37) and Eq.(39) together. The probability constraint between domains can effectively improve the semantic discrimination. Actually, l_1-norm can also be imposed on the difference between the probabilities of source and target samples.

6.3 Autoencoder-Based

As mentioned above, the training of DNNs needs a large amount of labeled source data. For unsupervised feature learning in domain adaptation, deep autoencoder based network adaptation framework was presented [215], [216], [217], [166], [218]. Generic auto-encoders are comprised of an encoder function $f(\cdot)$ and a decoder function $g(\cdot)$, which are typically trained to minimize the reconstruction error. Denoising autoencoders (DAE) were generally constructed with one-layer neural networks for reconstructing original data from partially or randomly corrupted data [221]. The denoising autoencoders can be stacked into a deep network (i.e., SDA), optimized by greedy layer-wise fashion based on stochastic gradient descent (SGD). The rational behind deep autoencoder based network adaptation is that the source data trained parameters of encoder and decoder can be adapted to represent those samples from a target domain.

In [215], Glorot et al. proposed a SDA based feature representation in conjunction with SVMs for sentiment analysis across different domains. Chen et al. [216] proposed a marginalized stacked denoising autoencoder (mSDA), which addressed two crucial limitations of SDAs, such as high computational cost and low scalability to high-dimensional features, by inducing a closed-form solution of parameters without SGD. In [217], Zhuang et al. proposed a supervised deep autoencoder for learning domain invariant features. The encoder is constructed with two encoding layers: embedding layer for domain disparity minimization and label encoding layer for softmax guided source classifier training. Suppose x, z and $\hat{x}$ to be the input sample, intermediate representation (encoded) and reconstructed output (decoded), respectively, then there is

$$z = f(x), \hat{x} = g(z) \tag{40}$$

where z is the intermediate feature representation of sample x. Generally, stacked deep autoencoder based TAL framework can be written as

$$\min_{f,g,\theta} \mathcal{J}(x,\hat{x}) + \lambda\Omega(z_s,z_t) + \beta\mathcal{L}(z_s,y_s,\theta) + \gamma\mathcal{R}(f,g) \tag{41}$$

where f is domain shared encoder, g is domain shared decoder, $\mathcal{J}(\cdot) = \mathcal{J}_s(\cdot) + \mathcal{J}_t(\cdot)$ represents the reconstruction error loss (e.g., l_2-norm squared loss), $\Omega(\cdot)$ is the distribution discrepancy metric between source feature z_s and target feature z_t, $\mathcal{L}(\cdot)$ is the classifier loss (e.g. cross-entropy) with parameter θ learned on the set(z_s, y_s), and $\mathcal{R}(\cdot)$ is the regularizer of the network parameters of f and g. In [217], Kullback-Leibler (KL) divergence [222] based distribution distance metric was considered. KL is a non-symmetric measure of the divergence between two probability distributions P and Q, which was defined as $D_{KL}(P||Q) = \sum_i P(i)\ln(\frac{P(i)}{Q(i)})$. Smaller value of $D_{KL}(\cdot)$ means higher similarity of two distributions. Due to that $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, a symmetric KL version was used in [217], in which the $\Omega(\cdot)$ in Eq.(41) was written as

$$\Omega(z_s,z_t) = D_{KL}(P_s||P_t) + D_{KL}(P_t||P_s) \tag{42}$$

where $P_s = \frac{\bar{z}_s}{\sum\bar{z}_s}$ and $P_t = \frac{\bar{z}_t}{\sum\bar{z}_t}$ represent the probability distribution of source and target domains. $\bar{z}_s$ and $\bar{z}_t$ represent the mean vector of encoded feature representations of source and target samples, respectively.

Similar to the reconstruction protocol in stacked autoencoder, a related work with deep reconstruction based on convolutional neural networks can be referred to as [219], in which the encoded source feature representation is feeded into the source classifier for visual classification and simultaneously into the decoder module for reconstructing the target data. Under this framework, a shared encoder for both domains can be learned.

6.4 Discussion and Summary

In this section, deep network adaptation advances are presented and categorized, which mainly contains three types of technical challenges: marginal alignment based, conditional alignment based and autoencoder based. A common characteristic of these methods is that the softmax guided cross-entropy loss based on labeled source data was minimized for classifier learning. In marginal alignment based models, the distribution discrepancy of feature representation from top layers is generally characterized by MMD. Besides that, the semantic similarity across domains was further characterized in conditional alignment based models. Different from both marginal and conditional alignment models, the autoencoder based ones tend to learn domain invariant feature embedding by imposing a Kullback-Leibler divergence in feature embedding layer.

Despite recent advances deep network adaptation faces several challenges: 1) a number of labeled source data is needed for training (fine-tuning) a deep network; 2) the confidence of an unlabeled target sample predicted to class k is sometimes very low when domain disparity is very large.

7 Adversarial Adaptation

Adversarial learning, originated from the generative adversarial net (GAN) [33], is a promising approach for generating pixel-level target samples or feature-level target representations by training robust DNNs. Currently, adversarial learning has become an increasing popular idea for addressing TAL issues, by minimizing the between-domain discrepancy through an adversarial objective (e.g., binary domain discriminator), instead of the generic MMD-based domain disparity in RKHS spaces. In fact, minimization of the domain disparity is amount to domain confusion in a learned feature space, where the domain discriminator cannot discriminate which domain a sample comes from. In this paper, the adversarial adaptation based TAL approaches are divided into three types: (i) *Gradient Reversal-Based*, (ii) *Minimax Optimization-Based* and (iii) *Generative Adversarial Net-Based*. The first two resort to feature-level domain confusion supervised by a domain discriminator for domain distribution discrepancy minimization, while the last one tends to pixel-level domain transfer by synthesizing implausible target domain images. The taxonomy of adversarial adaptation challenges is summarized in Table 5.

7.1 Gradient Reversal-Based

In adversarial optimization of DNNs between the general cross-entropy loss for source classifier learning and the domain discriminator for domain label prediction, Ganin and Lempitsky [223] firstly demonstrated that the domain adaptation behavior can be achieved by adding a simple but effective *gradient reversal layer (GRL)*. The augmented deep architecture can still be trained using standard stochastic gradient descent (SGD) based backpropagation. The gradient reversal based adversarial adaptation network consists of three parts: domain-invariant feature representation θ_f,

TABLE 5
Our Taxonomy of Adversarial Adaptation Challenges

ADVERSARIAL ADAPTATION	MODEL BASIS	REFERENCE
Gradient Reversal	Domain Confusion &GRL	[223], [224], [225], [226], [227], [228], [229], [230], [231]
Minimax Optimization	Domain Confusion &Game	[232], [233], [234] [235], [236], [237], [238], [239], [240]
GANs-Based	Pixel-level Image Synthesis	[241], [242], [243], [244], [245], [41], [246], [247], [248] [249], [48], [47]

visual classifier θ_c and domain classifier θ_d. Objectively, θ_f can be learned by trying to *minimize* the visual classifier loss L_c and simultaneously *maximize* the domain classifier loss L_d, such that the feature representation can be domain invariant (i.e. domain confusion) and class discriminative. Therefore, in backpropagation optimization of θ_f, the contributed gradients from losses L_c and L_d are $\frac{\partial L_c}{\partial \theta_f}$ and $-\lambda\frac{\partial L_d}{\partial \theta_f}$, respectively. The essence of GRL lies in the reversal gradient with negative multiplier $-\lambda I$.

More recently, the gradient reversal based adversarial strategy has been used for domain adaptation [224], [225], [226], [227] under CNN architecture, domain adaptive object detection [228] under Faster-RCNN framework, large-scale kinship verification [229], [230] and fine-grained visual classification [231] under Siamese network. By following a similar protocol with [223], in [224], [228], a domain classifier was designed as an adversarial objective for learning domain-invariant features by deploying a GRL layer. In [229], [230], two methods, AdvNet and Adv-Kin were proposed, in which a general Siamese network was constructed with three fully-connected (*fc*-) layers for similarity learning. The reversal gradient with negative multiplier $-\lambda I$ was placed in the 1^{st} *fc*-layer (MMD-loss), the generic contrastive loss was deployed in the 2^{nd} *fc*-layer and the softmax guided cross-entropy loss was deployed in the last *fc*-layer. In [226], Pei et al. argued that single domain discriminator based adversarial adaptation only aligns the between-domain distribution without exploiting the multimode structures. Therefore, they proposed a multi-adversarial domain adaptation (MADA) method based on GRL with multiple class-wise domain discriminators for capturing multimode structures, such that fine-grained alignment of different distributions is enabled. Also, Zhang et al. [227] proposed a collaborative adversarial network (CAN) by designing multiple domain classifiers, one for each feature extraction block in CNN.

7.2 Minimax Optimization-Based

In GANs, the two key parts G and D are often placed with an adversarial state, and generally solved by using a minimax based gaming optimization method [33]. Therefore, the minimax optimization based adversarial adaptation can be implemented for domain confusion, through an adversarial objective of the domain classifier or regressor [232], [233], [234], [235], [236], [237], [238], [239]. Minimax optimization based adversarial adaptation training of DNNs originated in 2015 [232], [233]. Domain confusion maximization based adversarial domain adaptation was first proposed by Tzeng et al. [232], in which an adversarial CNN framework was deployed with classification loss, softlabel loss and two adversarial objectives i.e., domain confusion loss and domain classifier loss. In [233], Ajakan et al. firstly proposed an adversarial training of stacked autoencoders (DANN) deployed with classification loss and an adversarial objective i.e., domain regressor loss.

Suppose the labeled source data trained visual classifier to be C, the domain discriminator to be D, and the feature representation to be F. The corresponding parameters are defined as θ_C, θ_D and θ_F. The general adversarial adaptation model aims to minimize the visual classifier loss $\mathcal{L}_C$ and maximize the domain discriminator loss $\mathcal{L}_D$ by learning θ_F, such that the feature representation function F can be more discriminative and domain-invariant. Simultaneously, the adversarial training aims to minimize the domain discriminator loss $\mathcal{L}_D$ under θ_F. Generally, maximizing $\mathcal{L}_D$ is amount to maximizing the domain confusion, such that it cannot discriminate which domain the samples come from, and vice versa. The above process can be generally formulated as the following adversarial adaptation model,

$$\begin{gathered}\min_{\theta_C,\theta_F} \mathcal{L}_C(\mathcal{D}_S,\mathcal{Y}_S;\theta_C,\theta_F) - \lambda\mathcal{L}_D(\mathcal{D}_S,\mathcal{D}_T,\theta_D;\theta_F) \\ \min_{\theta_D} \mathcal{L}_D(\mathcal{D}_S,\mathcal{D}_T,\theta_F;\theta_D)\end{gathered} \tag{43}$$

where $\mathcal{D}_S$ and $\mathcal{D}_T$ mean the source and target domain samples, $\mathcal{Y}_S$ denotes the source data labels.

Under this basic framework in Eq.(43), Tzeng et al. [235] further proposed an adversarial discriminative domain adaptation (ADDA) method, in which two CNNs were separately learned for source and target domain. The training of source CNN relied only on the source data and labels by minimizing the cross-entropy loss $\mathcal{L}_C$, while the target CNN and the domain discriminator loss $\mathcal{L}_D$ was alternatively trained in an adversarial fashion with the source CNN fixed. Rozantsev et al. [237] proposed a residual parameter transfer model with adversarial domain confusion supervised by a domain classifier, in which the residual transform between domains was deployed in convolutional layers. For augmenting the domain-specific feature representation, Long et al. [238] proposed a conditional domain adversarial network (CDAN), in which the feature representation and classifier prediction were integrated via multilinear map for jointly learning the domain classifier. More recently, Saito et al. [240] proposed a novel adversarial strategy, i.e., maximum classifier discrepancy (MCD), which aims to maximize the discrepancy between two classifiers' outputs instead of domain discriminator. The feature extractor aims to minimize the two classifiers' discrepancy. They argued that the general domain discriminator does not take into account the task-specific decision boundaries between classes, which may lead to ambiguous features near class boundaries from the feature extractor.

7.3 Generative Adversarial Net-Based

In generative adversarial net (GAN) [33] and its variants, two key parts: generator G and discriminator D are generally composed. The generator G aims to synthesize implausible images by using the encoder and decoder, while the discriminator D plays a role in identification of authenticity by recognizing a sample to be true or false. A minimax gaming based alternative optimization scheme is generally used for solving G and D. In TAL studies, started from 2017, GAN based models have been presented to synthesize distribution approximated pixel-level images with target domain and then enable the cross-domain image classification by using synthesized image samples (e.g., objects, scenes, pedestrians and faces, etc.) [241], [242], [243], [244], [245], [41], [246], [247], [248].

Under the CycleGAN framework proposed by Zhu et al. [40], Hoffman et al. [241] firstly proposed a cycle-consistent adversarial domain adaptation model (CyCADA) for adapting representations in both pixel-level and feature-level without requiring aligned pairs, by jointly minimizing pixel loss, feature loss, semantic loss and cycle consistence loss. Bousmalis et al. [242] and Taigman et al. [243] proposed GAN-based models for unsupervised image-level domain adaptation, which aims to adapt source domain images to appear as if drawn from target domain with well-preserved identity. In [244], Hu et al. proposed a duplex GAN (DupGAN) for image-level domain transformation, in which the duplex discriminators, one for each domain, were trained against the generator for ensuring the reality of the domain transformation. Murez et al. [245] and Hong et al. [246] proposed image-to-image translation based domain adaptation models by leveraging GAN and synthetic data for semantic segmentation of the target domain images. Person re-identification (ReID) is typical cross-domain feature match and retrieval problem [178], [164]. Recently, for addressing ReID challenges in complex scenarios, GAN-based domain adaptation was presented for implausible person image generation from source domain to target domain [249], [247], [248], across different visual cues and styles, such as poses, backgrounds, lightings, resolutions, seasons, etc. Additionally, GAN based cross-domain facial image generation for pose-invariant face representation, face frontalization and rotation were intensively studied [36], [50], [48], [47], all of which tend to address domain adaptation and transfer problems in face recognition across poses.

7.4 Discussion and Summary

In this section, adversarial adaptation is presented with three streams, including gradient reversal, minimax optimization and generative adversarial net (GAN). The gradient reversal and minimax optimization share a common characteristic, i.e., feature-level adaptation, by introducing a domain discriminator based adversarial objective for training against the feature extractor. The difference between them is the *against strategy*. Different from both of them, GAN-based adversarial adaptation focuses on pixel-level adaptation, i.e., image generation from source domain to a target, such that the synthesized implausible images are as if drawn from target domain.

Adversarial adaptation is recognized to be an emerging perspective, despite these advances it still faces with several challenges: 1) the domain discriminator is easily overtrained; 2) maximizing only domain confusion easily leads to class bias; 3) the gaming between feature generator and discriminator is human dependent.

8 Benchmark Datasets

In this section, the benchmark datasets for testing TAL models are introduced to facilitate readers' impression on how to start studies of transfer adaptation learning. Totally, 12 benchmark datasets including Office-31 (3DA) [5], Office+Caltech-10 (4DA) [5], [112], [203], [250], MNIST+USPS [130], [131], Multi-PIE [130], [131], COIL-20 [251], MSRC+VOC2007 [116], IVLSC [252], [253], AwA [254], Cross-dataset Testbed [1], Office Home [255], ImageCLEF [256], and P-A-C-S [252] are summarized, each of which contains at least 2 different domains.

8.1 Office-31 (3DA)

Office-31 is a popular benchmark for visual domain transfer, which includes 31 categories of samples drawn from three different domains, i.e., Amazon (A), DSLR (D) and Webcam (W). Amazon consists of online e-commerce pictures, DSLR contains high-resolution pictures and Webcam contains low-resolution pictures taken by a web camera. There are totally 4652 images, composed of 2817, 498 and 795 images from domain A, D and W, respectively. In feature extraction, (1) for shallow features, 800-dimensional feature vectors extracted by the Speed Up Robust Features (SURF) were used, and (2) for deep features, 4096-dimensional feature vectors extracted from pre-trained AlexNet or VGG-net were used. In model evaluation, six kinds of source-target domain pairs were tested, i.e., A→D, A→W, D→A, D→W , W→A, W→D.

8.2 Office+Caltech-10 (4DA)

This 4DA dataset contains 4 domains, in which 3 domains (A, D, W) are from the Office-31 and another domain (C) is from Caltech-256, a benchmark containing 30,607 images of 256 classes in object recognition. The common 10 classes among the Office-31 and Caltech-256 were selected to form the 4DA, and therefore 2,533 images composed of 958, 157, 295 and 1123 images from domain A, D, W and C were collected. In evaluation, 12 tasks with different source-target domain pairs are addressed, i.e., A→D, A→C, A→W, D→A, D→C, D→W, C→A, C→D, C→W, W→A, W→C, W→D.

8.3 MNIST+USPS

MNIST and USPS are two benchmarks containing 10 categories of digit images under different distribution for handwritten digit recognition, and therefore qualified for TAL tasks. The MNIST includes 60,000 training pictures and 10,000 test pictures. The USPS includes 7291 training pictures and 2007 test pictures. For TAL tasks, 2000 pictures and 1800 pictures were randomly selected from MNIST and USPS, respectively. For feature extraction, each image was resized into 16×16 and a 256-dimensional feature vector that encode the pixel values was finally extracted. In evaluation, 2 cross-domain tasks, i.e., MNIST→USPS and USPS→MNIST are addressed.

8.4 Multi-PIE

Multi-PIE is a benchmark with poses, illuminations and expressions in face recognition, which includes 41,368 faces of 68 different identities. For TAL tasks, (1) face recognition across poses is generally evaluated on five different face orientations, including C05: left pose, C07: upward pose, C09: downward pose, C27: front pose and C29: right pose. Totally, 3332, 1629, 1632, 3329, and 1632 facial images are contained in C05, C07, C09, C27 and C29. Therefore, 20 tasks were evaluated, i.e., C05→C07, C05→C09, C05→C27, etc.; (2) face recognition across illuminations and exposure conditions is evaluated by randomly selecting two sets: PIE1 and PIE2 from front face images. Two tasks: PIE→PIE2 and PIE2→PIE1 were evaluated.

8.5 COIL-20

COIL-20 is a 3D object recognition benchmark containing 1440 images of 20 object categories. By rotating each object class horizontally of 5 degrees, 72 images per class after rotating 360 degrees were obtained. For TAL tasks, two disjoint subsets with different distribution i.e., COIL1 and COIL2 were prepared, where COIL1 contains the images in [0°, 85°] U [180°, 265°] and the images of COIL2 are in [90°, 175°] U [270°, 355°]. Therefore, two cross-domain tasks i.e., COIL1→COIL2 and COIL2→COIL1 were evaluated.

8.6 MSRC+VOC2007

The MSRC contains 4323 images of 18 categories and VOC2007 contains 5011 images of 20 categories. 1269 and 1530 images w.r.t six common categories, i.e., *aeroplane, bicycle, bird, car, cow* and *sheep*, were finally selected from MSRC and VOC2007, respectively. In feature representation, 128-dimensional DenseSIFT features were extracted for cross-domain image classification tasks, i.e., MSRC→VOC2007 and VOC2007→MSRC.

8.7 IVLSC

IVLSC is a large-scale image dataset containing five subsets, i.e., ImageNet (I), VOC2007 (V), LabelMe (L), SUN09 (S), and Caltech (C). For TAL tasks, 7341, 3376, 2656, 3282, and 1415 samples w.r.t. five common categories i.e., *bird, cat, chair, dog* and *human*, were randomly selected from I, V, L, S, and C domains, respectively. In feature representation, 4096-dimensional DeCaf6 deep features were extracted for cross-domain image classification under 20 tasks, i.e., I→V, I→L, I→S, I→C, ..., C→I, C→V, C→L, C→S.

8.8 AwA

AwA is an animal identification dataset containing 30,475 images of 50 categories, which provides a benchmark due to the inherent data distribution difference. This data set is currently less used in evaluating TAL algorithms.

8.9 Cross-dataset Testbed

This benchmark contains 10,473 images of 40 categories, collected from three domains: 3,847 images in Caltech256 (C), 4,000 images in ImageNet (I), and 2,626 images in SUN (S). In feature extraction, the 4096-dimensional DeCAF7 deep features were used for cross-domain image classification tasks, i.e., C→I, C→S, I→C, I→S, S→C, S→I.

8.10 Office Home

Office Home is a relatively new benchmark containing 15,585 images of 65 categories, collected from 4 domains, i.e., (1) Art (Ar): artistic depictions of objects in the form of sketches, paintings, ornamentation, etc.; (2) Clipart (Cl): collection of clipart images; (3) Product (Pr): images of objects without a background, akin to the Amazon category in Office dataset; (4) Real-World (RW): images of objects captured with a regular camera. In detail, there contains 2421, 4379, 4428 and 4357 images in *Ar, Cl, Pr* and *RW* domains, respectively. In evaluation, 12 cross-domain tasks were tested, e.g., Ar→Cl, Ar→Pr, Ar→RW, Cl→Ar, etc.

8.11 ImageCLEF

This benchmark includes 1800 images of 12 categories, which were drawn from 3 domains: 600 images in Caltech 256 (C), 600 images in ImageNet ILSVRC2012 (I), and 600 images in Pascal VOC2012 (P). Therefore, 6 cross-domain tasks i.e., C→I, C→P, I→C, I→P, P→C, P→I were evaluated.

8.12 P-A-C-S

PACS is a new benchmark containing 7 common categories: *dog, elephant, giraffe, guitar, horse, house* and *person*, from 4 domains, i.e., 1670 images in Photo (P), 2048 images in Art Painting (A), 2344 images in Cartoon (C), and 3929 images in Sketch (S). In feature representation, 4096-dimensional VGG-M deep features were used and 12 cross-domain tasks are evaluated, e.g., P→A, P →C, P→S, A→P, A→C, etc.

8.13 Discussion and Summary

In this section, 12 benchmarks constructed based on popular datasets in computer vision such as ImageNet, ILSVRC, PASCAL VOC, Caltech-256, multi-PIE and MNIST for addressing cross-domain image classification tasks are presented. Despite these endeavors made by researchers, more benchmarks in cross-domain vision understanding problems we could see, namely: object detection, semantic segmentation, visual relation modeling, scene parsing, etc. are future challenges for transfer adaptation learning.

9 Conclusion

Transfer adaptation learning is an energetic research field which aims to learn domain adaptive representations and classifiers from source domains toward representing and recognizing samples from a distribution different but semantic related target domain. This paper surveyed recent advances in transfer adaptation learning in the past decade and present a new taxonomy of five technical challenges being faced by researchers: instance re-weighting adaptation, feature adaptation, classifier adaptation, deep network adaptation and adversarial adaptation. Besides, 12 visual benchmarks that address multiple cross-domain recognition tasks are collected and summarized to help facilitate researchers' insight on the tasks and scenarios that transfer adaptation learning aims to address.

The proposed taxonomy of transfer adaptation learning challenges in this paper provides a framework for researchers better understanding and identifying the status of the field, future research challenges and directions. Each challenge was summarized with a discussion of existing problems and future direction, which, we believe, are worth studying for better capturing general domain knowledge, toward universal machine learning. Throughout the entire research lines, one specific research area of transfer adaptation learning that seems to be still under-studied is the co-adaptation of multiple but heterogeneous domains, which goes beyond two homogeneous domains. This challenge is more approaching real-world scenarios that numerous domains can be found, and co-adaptation expects to capture commonality and specificity among multiple domains. Additionally, an open question that remains to be unanswered is when will we need transfer adaptation learning for a given application scenario? the basic analyzing condition of whether *cross-domain* happens is still not clear. We observe these promising directions of transfer adaptation learning for future research.

Acknowledgment

The author would like to thank the pioneer researchers in transfer learning, domain adaptation and other related fields. The author would also like to thank Dr. Mingsheng Long and Dr. Lixin Duan for their kindly help in providing insightful discussions.

References

[1] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR*, 2011, pp. 1521–1528.

[2] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Stat. Plan. Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[3] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *ACM MM*, 2007, pp. 188–197.

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

[5] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *ECCV*, 2010.

[6] D. Cook, K. Feuz, and N. Krishnan, "Transfer learning for activity recognition: a survey," *Knowl. Inf. Syst.*, vol. 36, pp. 537–556, 2013.

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

[8] V. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation," *IEEE Signal Processing Magazine*, pp. 53–69, 2015.

[9] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1019–1034, 2015.

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

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

[12] K. Weiss, T. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, pp. 1–40, 2016.

[13] S. Salaken, A. Khosravi, T. Nguyen, and S. Nahavandi, "Extreme learning machine based transfer learning algorithms: A survey," *Neurocomputing*, vol. 267, pp. 516–524, 2017.

[14] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[16] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2015, pp. 770–778.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[19] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region based fully convolutional networks," in *NIPS*, 2016.

[20] W. Liu, D. Anguelov, D. Erhan, S. Christian, S. Reed, C. Fu, and A. Berg, "Ssd: single shot multibox detector," in *ECCV*, 2016.

[21] C. Fu, W. Liu, A. Ranga, A. Tyagi, and A. Berg, "Dssd: Deconvolutional single shot detector," in *arXiv*, 2017.

[22] E. Ahmed, M. Jones, and T. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015.

[23] D. Chung, K. Tahboub, and E. Delp, "A two steam siamese convolutional neural network for person re-identification," in *ICCV*, 2017.

[24] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *CVPR*, 2016.

[25] L. Hou, D. Samaras, T. Kurc, Y. Gao, J. Davis, and J. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *CVPR*, 2016.

[26] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.

[27] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.

[28] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer learning from deep features for remote sensing and poverty mapping," in *AAAI*, 2016.

[29] N. Jean, M. Burke, M. Xie, W. Davis, D. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, p. 790794, 2016.

[30] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[31] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 645–657, 2017.

[32] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," *JMLR*, vol. 27, p. 1737, 2012.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *arXive*, 2014.

[34] M. Mirza and S. Osindero, "Conditional generative adversarial nets," in *arXiv*, 2014.

[35] L. Tran, X. Yin, and X. Liu, "Representation learning by rotating your faces," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2018.

[36] ——, "Disentangled representation learning gan for pose-invariant face recognition," in *CVPR*, 2017.

[37] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *ICML*, 2016.

[38] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *ICLR*, 2016.

[39] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134.

[40] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[41] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018, pp. 8789–8797.

[42] M. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *NIPS*, 2017.

[43] D. Yoo, N. Kim, S. Park, A. Paek, and I. Kweon, "Pixel-level domain transfer," in *ECCV*, 2016.

[44] L. Gatys, A. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *CVPR*, 2016, pp. 2414–2423.

[45] L. Gatys, A. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, "Controlling perceptual factors in neural style transfer," in *CVPR*, 2017, pp. 3985–3993.

[46] J. Johnson, A. Alahi, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016, pp. 694–711.

[47] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *CVPR*, 2018.

[48] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *ICCV*, 2017.

[49] J. Cao, Y. Hu, B. Yu, R. He, and Z. Sun, "Load balanced gans for multi-view face image synthesis," in *arXiv:1802.07447*, 2018.

[50] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *ICCV*, 2017.

[51] H. Yang, D. Huang, Y. Wang, and A. Jain, "Learning face age progression: A pyramid architecture of gans," in *CVPR*, 2018.

[52] A. Evgeniou and M. Pontil, "Multi-task feature learning," in *NIPS*, 2007.

[53] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci*, vol. 55, no. 1, pp. 119–139, 1997.

[54] T. Dietterich, "Machine learning: Four current directions," *AI Mag.*, vol. 18, no. 4, pp. 97–136, 1997.

[55] X. Wu, V. Kumar, J. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z. Zhou, M. Steinbach, D. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, pp. 1–37, 2008.

[56] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 1, pp. 1–10, 2017.

[57] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-Supervised Learning*, 2006.

[58] Z. Zhou and M. Li, "Semi-supervised learning by disagreement," *Knowledge and Information Systems*, vol. 24, no. 3, pp. 415–439, 2010.

[59] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine Learning*, vol. 39, pp. 103–134, 2000.

[60] D. Miller and H. Uyar, "A mixture of experts classifier with learning based on both labeled and unlabeled data," in *NIPS*, 1997, pp. 571–577.

[61] D. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *NIPS*, 2014.

[62] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *AISTATS*, 2005, pp. 57–64.

[63] T. Joachims, "Transductive inference for text classication using support vector machines," in *ICML*, 1999, pp. 200–209.

[64] Y. Li, I. Tsang, J. Kwok, and Z. Zhou, "Convex and scalable weakly labeled svms," *Journal of Machine Learning Research*, vol. 14, pp. 2151–2188, 2013.

[65] Z. Zhou and M. Li, "Tri-training: exploiting unlabeled data using three classifiers," *IEEE Trans. Knowledge Data Engineering*, vol. 17, pp. 1529–1541, 2005.

[66] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *The 11th Int' Conf' Computational Learning Theory*, 1998, pp. 92–100.

[67] Z. Zhou and M. Li, "Semi-supervised learning by disagreement," *Knowledge and Information Systems*, vol. 24, pp. 415–439, 2010.

[68] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *ICML*, 2001, pp. 19–26.

[69] M. Belkin and P. Niyogi, "Semi-supervised learning on riemannian manifolds," *Machine Learning*, vol. 56, no. 1-3, pp. 209–239, 2004.

[70] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[71] Z. Yang, W. Cohen, and R. Salakhutdinov, "Revisiting semi-supervised learning with graph embeddings," in *ICML*, 2016.

[72] X. Zhu, "Semi-supervised learning literature survey," *Technical Report 1530*, 2008.

[73] I. Triguero, S. Garcia, and F. Herrera, "Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study," *Knowledge and Information Systems*, vol. 42, no. 2, pp. 245–284, 2015.

[74] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," *Acm Sigir Forum*, vol. 29, no. 2, pp. 13–19, 1994.

[75] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification." in *ICML*, 2000, pp. 999–1006.

[76] E. Elhamifar, G. Sapiro, A. Yang, and S. S. Sasrty, "A convex optimization framework for active learning," in *ICCV*, 2013, pp. 209–216.

[77] B. Settles, "Active learning literature survey," in *Technical Report*, 2010.

[78] M. Rohrbach, M. Stark, and B. Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in *CVPR*, 2011, pp. 1641–1648.

[79] C. Lampert, H. Nickishch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[80] Y. Fu, T. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2332–2345, 2015.

[81] Z. Ding, M. Shao, and Y. Fu, "Generative zero-shot learning via low-rank embedded semantic dictionary," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2018.

[82] L. Niu, J. Cai, A. Veeraraghavan, and L. Zhang, "Zero-shot learning via category-specific visual-semantic mapping and label refinement," *IEEE Trans. Image Processing*, 2018.

[83] S. Rahman, S. Khan, and F. Porikli, "A unified approach for conventional zero-shot, generalized zero-shot and few-shot learning," *IEEE Trans. Image Processing*, pp. 1–16, 2018.

[84] Y. Yu, Z. Ji, J. Guo, and Y. Pang, "Transductive zero-shot learning with adaptive structural embedding," *IEEE Trans. Neural Networks and Learning Systems*, vol. 29, no. 9, pp. 4116–4127, 2018.

[85] Y. Yu, Z. Ji, X. Li, J. Guo, Z. Zhang, H. Ling, and F. Wu, "Transductive zero-shot learning with self-training dictionary approach," *IEEE Trans. Cybernetics*, vol. 48, no. 10, pp. 2908–2919, 2018.

[86] X. Xu, T. Hospedales, and S. Gong, "Transductive zero-shot action recognition by word-vector embedding," *International Journal of Computer Vision*, vol. 123, no. 3, pp. 309–333, 2017.

[87] X. Li, Y. Guo, and D. Schuurmans, "Semi-supervsied zero-shot classification with label representation learning," in *ICCV*, 2015, pp. 4211–4219.

[88] J. Song, C. Shen, Y. Yang, Y. Liu, and M. Song, "Transductive unbiased embedding for zero-shot learning," in *CVPR*, 2018, pp. 1024–1033.

[89] W. Scheirer, A. Rocha, A. Sapkota, and T. Boult, "Towards open set recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2013.

[90] W. Scheirer, L. Jain, and T. Boult, "Probability models for open set recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 11, pp. 2317–2324, 2014.

[91] F. Li and H. Wechsler, "Open set face recognition using transduction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1686–1697, 2005.

[92] H. Zhang and V. Patel, "Sparse representation-based open set recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1690–1696, 2016.

[93] P. Panareda Busto and J. Gall, "Open set domain adaptation," in *ICCV*, 2017, pp. 754–763.

[94] P. Panareda Busto, A. Iqbal, and J. Gall, "Open set domain adaptation for image and action recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2018.

[95] V. Kumar Verma, G. Arora, A. Mishra, and P. Rai, "Generalized zero-shot learning via synthesized examples," in *CVPR*, 2018, pp. 4281–4289.

[96] J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Scholkopf, "Correcting sample selection bias by unlabeled data," in *NIPS*, 2007, pp. 1–8.

[97] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *ICML*, 2004.

[98] J. Jiang and C. Zhai, "Instance weighting for domain adaptation in nlp," in *ACL*, 2007, pp. 264–271.

[99] R. Wang, M. Utiyama, L. Liu, K. Chen, and E. Sumita, "Instance weighting for neural machine translation domain adaptation," in *ACL*, 2017, pp. 1482–1488.

[100] W. Dai, Q. Yang, G. R. Xue, and Y. Yu, "Boosting for transfer learning," in *ICML*, 2007, pp. 193–200.

[101] S. Chen, F. Zhou, and Q. Liao, "Visual domain adaptation using weighted subspace alignment," in *Visual Communications and Image Processing*, 2017.

[102] W. S. Chu, l. T. F. De, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," 2013, pp. 3515–3522.

[103] 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 *CVPR*, 2017, pp. 2272–2281.

[104] E. H. Zhong, W. Fan, J. Peng, K. Zhang, J. Ren, D. S. Turaga, and O. Verscheure, "Cross domain distribution adaptation via kernel mapping," in *ACM SIGKDD*, 2009, pp. 1027–1036.

[105] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *CVPR*, 2014, pp. 1410–1417.

[106] M. Chen, K. Q. Weinberger, and J. C. Blitzer, "Co-training for domain adaptation," in *NIPS*, 2011.

[107] Q. Chen, Y. Liu, Z. Wang, I. Wassell, and K. Chetty, "Re-weighted adversarial adaptation network for unsupervised domain adaptation," in *CVPR*, 2018, pp. 7976–7985.

[108] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *ICCV*, 2013, pp. 2960–2967.

[109] A. Gretton, K. Borgwardt, M. Rasch, B. Schoelkopf, and A. Smola, "A kernel method for the two-sample-problem," in *NIPS*, 2006.

[110] A. Gretton, K. Borgwardt, M. Rasch, B. Scholkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, pp. 723–773, 2012.

[111] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *ICCV*, 2011, pp. 999–1006.

[112] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *CVPR*, 2012, pp. 2066–2073.

[113] B. Sun and K. Saenko, "Subspace distribution alignment for unsupervised domain adaptation," in *BMVC*, 2015, pp. 24.1–24.10.

[114] J. Liu and L. Zhang, "Optimal projection guided transfer hashing for image retrieval," in *AAAI*, 2018.

[115] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Networks*, vol. 22, no. 2, p. 199, 2011.

[116] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *ICCV*, 2014, pp. 2200–2207.

[117] M. Xiao and Y. Guo, "Feature space independent semi-supervised domain adaptation via kernel matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 54–66, 2015.

[118] S. Si, D. Tao, and B. Geng, "Bregman divergence-based regularization for transfer subspace learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 22, no. 7, pp. 929–942, 2010.

[119] Z. Ding and Y. Fu, "Robust transfer metric learning for image classification," *IEEE Trans. Image Processing*, vol. 26, no. 2, p. 660670, 2017.

[120] H. Wang, W. Wang, C. Zhang, and F. Xu, "Cross-domain metric learning based on information theory," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, p. 20992105.

[121] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *ECCVW*, 2016, pp. 443–450.

[122] S. Herath, M. Harandi, and F. Porikli, "Learning an invariant hilbert space for domain adaptation," in *CVPR*, 2017, pp. 3845–3854.

[123] H. Daume III, A. Kumar, and A. Saha, "Frustratingly easy semi-supervised domain adaptation," in *ACL*, 2010, pp. 53–59.

[124] W. Li, L. Duan, D. Xu, and I. Tsang, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1134–1148, 2014.

[125] R. Volpi, P. Morerio, S. Savarese, and V. Murino, "Adversarial feature augmentation for unsupervised domain adaptation," in *CVPR*, 2018, pp. 5495–5504.

[126] L. Zhang, S. Wang, G. Huang, W. Zuo, J. Yang, and D. Zhang, "Manifold criterion guided transfer learning via intermediate domain generalization," *IEEE Trans. Neural Networks and Learning Systems*, 2019.

[127] I. H. Jhuo, D. Liu, D. T. Lee, and S. F. Chang, "Robust visual domain adaptation with low-rank reconstruction," in *CVPR*, 2012, pp. 2168–2175.

[128] M. Shao, D. Kit, and Y. Fu, "Generalized transfer subspace learning through low-rank constraint," *International Journal of Computer Vision*, vol. 109, no. 1-2, pp. 74–93, 2014.

[129] J. Fu, L. Zhang, B. Zhang, and W. Jia, "Guided learning: A new paradigm for multi-task classification," in *CCBR*, 2018, pp. 239–246.

[130] L. Zhang, W. Zuo, and D. Zhang, "Lsdt: Latent sparse domain transfer learning for visual adaptation," *IEEE Trans. Image Processing*, vol. 25, no. 3, pp. 1177–1191, 2016.

[131] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, "Discriminative transfer subspace learning via low-rank and sparse representation." *IEEE Trans. Image Processing*, vol. 25, no. 2, pp. 850–863, 2016.

[132] L. Zhang, J. Yang, and D. Zhang, "Domain class consistency based transfer learning for image classification across domains," *Information Sciences*, vol. 418-419, pp. 242–257, 2017.

[133] S. Wang, L. Zhang, and W. Zuo, "Class-specific reconstruction transfer learning via sparse low-rank constraint," in *ICCVW*, 2017, pp. 949–957.

[134] S. Shekhar, V. Patel, H. Nguyen, and R. Chellappa, "Generalized domain-adaptive dictionaries," in *CVPR*, 2013, pp. 361–368.

[135] H. Xu, J. Zheng, and R. Chellappa, "Bridging the domain shit by domain adaptive dictionary learning," in *BMVC*, 2015.

[136] S. Shekhar, V. Patel, H. Van Nguyen, and R. Chellappa, "Coupled projections for adaptation of dictionaries," *IEEE Trans. Image Processing*, vol. 24, no. 10, pp. 2941–2954, 2015.

[137] Q. Qiu and R. Chellappa, "Compositional dictionaries for domain adaptive face recognition," *IEEE Trans. Image Processing*, vol. 24, no. 12, pp. 5152–5165, 2015.

[138] Q. Qiu, V. Patel, P. Turaga, and R. Chellappa, "Domain adaptive dictionary learning," in *ECCV*, 2012.

[139] J. Ni, Q. Qiu, and R. Chellappa, "Subspace interpolation via dictionary learning for unsupervised domain adaptation," in *CVPR*, 2013, pp. 692–699.

[140] F. Zhu and L. Shao, "Weakly-supervised cross-domain dictionary learning for visual recognition," *International Journal of Computer Vision*, vol. 109, no. 1-2, pp. 42–59, 2014.

[141] B. Lu, R. Chellappa, and N. Nasrabadi, "Incremental dictionary learning for unsupervised domain adaptation," in *BMVC*, 2015.

[142] B. Kulis, K. Saenko, and T. Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *CVPR*, 2015, pp. 1785–1792.

[143] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1853–1865, 2017.

[144] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann, "Unsupervised domain adaptation by domain invariant projection," in *ICCV*, 2013, pp. 769–776.

[145] M. Long, J. Wang, J. Sun, and P. Yu, "Domain invariant transfer kernel learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 27, no. 6, pp. 1519–1532, 2015.

[146] L. Zhang, Y. Liu, Z. He, J. Liu, P. Deng, and X. Zhou, "Anti-drift in e-nose: A subspace projection approach with drift reduction," *Sensors and Actuators B: Chemical*, vol. 253, pp. 407–417, 2017.

[147] M. Ghifary, D. Balduzzi, W. Bastiaan Kleijn, and M. Zhang, "Scatter component analysis: A unified framework for domain adaptation and domain generalization," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 39, no. 7, pp. 1414–1430, 2017.

[148] J. Zhang, W. Li, and P. Ogunbona, "Joint geometrical and statistical alignment for visual domain adaptation," *CVPR*, pp. 5150–5158, 2017.

[149] L. Zhang, Y. Liu, and P. Deng, "Odor recognition in multiple e-nose systems with cross-domain discriminative subspace learning," *IEEE Trans. Instrumentation and Measurement*, vol. 66, no. 7, pp. 1679–1692, 2017.

[150] H. Lu, C. Shen, Z. Cao, Y. Xiao, and A. D. H. Van, "An embarrassingly simple approach to visual domain adaptation." *IEEE Trans. Image Processing*, 2018.

[151] S. Li, S. Song, and G. Huang, "Domain invariant and class discriminative feature learning for visual domain adaptation," *IEEE Trans. Image Processing*, vol. 27, no. 9, pp. 4260–4273, 2018.

[152] S. Li, S. Song, G. Huang, and C. Wu, "Cross-domain extreme learning machines for domain adaptation," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, 2018.

[153] A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *ALT*, 2005.

[154] K. Yan, L. Kou, and D. Zhang, "Learning domain-invariant subspace using domain features and independence maximization," *IEEE Trans. Cybernetics*, vol. 48, no. 1, pp. 288–299, 2017.

[155] Y. Zhang and D. Yeung, "Transfer metric learning by learning task relationships," in *ACM SIGKDD*, 2010, pp. 1199–1208.

[156] J. Hu, J. Lu, Y.-P. Tan, and J. Zhou, "Deep transfer metric learning," *IEEE Trans. Image Processing*, vol. 25, no. 12, 2016.

[157] B. Geng, D. Tao, and C. Xu, "Daml: Domain adaptation metric learning," *IEEE Trans. Image Processing*, vol. 20, no. 10, pp. 2980–2989, 2011.

[158] Y. Xu, S. Pan, H. Xiong, Q. Wu, R. Luo, H. Min, and H. Song, "A unified framework for metric transfer learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 29, no. 6, pp. 1158–408, 2017.

[159] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *AAAI*, 2016, pp. 153–171.

[160] Z. Zhang, M. Wang, Y. Huang, and A. Nehorai, "Aligning infinite-dimensional covariance matrices in reproducing kernel hilbert spaces for domain adaptation," in *CVPR*, 2018, pp. 3437–3445.

[161] L. Li and Z. Zhang, "Semi-supervised domain adaptation by covariance matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2018.

[162] H. Daume III, "Frustratingly easy domain adaptation," in *arXiv*, 2009.

[163] H. Daume III, A. Kumar, and A. Saha, "Co-regularization based semi-supervised domain adaptation," in *ACL*, 2010.

[164] Y. Chen, X. Zhu, W. Zheng, and J. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 392–408, 2018.

[165] X. Shi, Q. Liu, W. Fan, P. Yu, and R. Zhu, "Transfer learning on heterogenous feature spaces via spectral transformation," in *ICDM*, 2010.

[166] J. Zhou, S. Pan, I. Tsang, and Y. Yan, "Hybrid heterogeneous transfer learning through deep learning," in *AAAI*, 2014, pp. 2213–2219.

[167] Y. Yan, W. Li, H. Wu, H. Min, M. Tan, and Q. Wu, "Semi-supervised optimal transport for heterogeneous domain adaptation," in *IJCAI*, 2018.

[168] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *ICML*, 2010, pp. 663–670.

[169] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2012.

[170] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Eprint Arxiv*, vol. 9, 2013.

[171] J. Wright, A. Ganesh, S. Rao, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization," in *NIPS*, 2009.

[172] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *CVPR*, 2009, pp. 2790–2797.

[173] ——, "Sparse subspace clustering: Algorithms, theory, and applications," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

[174] M. Soltanolkotabi, E. Elhamifar, and E. Candes, "Robust subspace clustering," *Ann. Statist.*, vol. 42, no. 2, pp. 669–699, 2014.

[175] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[176] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, 2012.

[177] Z. Jiang, Z. Lin, and L. Davis, "Label consistent k-svd: Learning a discriminative dictionary for recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2651–2664, 2013.

[178] S. Li, M. Shao, and Y. Fu, "Cross-view projective dictionary learning for person re-identification," in *IJCAI*, 2015.

[179] L. Duan, D. Xu, and S. F. Chang, "Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach," in *CVPR*, 2012, pp. 1338–1345.

[180] L. Zhang and D. Zhang, "Domain adaptation extreme learning machines for drift compensation in e-nose systems," *IEEE Trans. Instrumentation and Measurement*, vol. 64, no. 7, pp. 1790–1801, 2015.

[181] A. Royer and C. Lampert, "Classifier adaptation at prediction time," in *CVPR*, 2015.

[182] L. Duan, I. Tsang, D. Xu, and S. Maybank, "Domain transfer svm for video concept detection," in *CVPR*, 2009.

[183] L. Duan, D. Xu, I. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," in *CVPR*, 2010.

[184] D. Lixin, X. Dong, T. Ivor Wai-Hung, and L. Jiebo, "Visual event recognition in videos by learning from web data," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1667–1680, 2012.

[185] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, p. 465, 2012.

[186] W. Li, Z. Xu, D. Xu, D. Dai, and L. Van Gool, "Domain generalization and adaptation using low rank exemplar svms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 40, no. 5, pp. 1114–1127, 2017.

[187] M. Long, J. Wang, G. Ding, S. Pan, and P. Yu, "Adaptation regularization: a general framework for transfer learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 26, no. 5, p. 10761089, 2014.

[188] L. Zhang and D. Zhang, "Robust visual knowledge transfer via extreme learning machine-based domain adaptation," *IEEE Trans. Image Processing*, vol. 25, no. 10, pp. 4959–4973, 2016.

[189] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, "Visual domain adaptation with manifold embedded distribution alignment," 2018.

[190] Y. Cao, M. Long, and J. Wang, "Unsupervised domain adaptation with distribution matching machines," in *AAAI*, 2018.

[191] T. Yao, Y. Pan, C. Ngo, H. Li, and T. Mei, "Semi-supervised domain adaptation with subspace learning for visual recognition," in *CVPR*, 2015, pp. 2142–2150.

[192] M. Gonen and A. Margolin, "Kernelized bayesian transfer learning," in *AAAI*, 2014, pp. 1831–1839.

[193] A. Ramachandran, S. Gupta, S. Rana, and S. Venkatesh, "Information-theoretic transfer learning framework for bayesian optimisation," in *ECML*, 2019, pp. 827–842.

[194] B. Liu and N. Vasconcelos, "Bayesian model adaptation for crowd counts," in *ICCV*, 2015, pp. 4175–4183.

[195] B. Gholami, O. Rudovic, and V. Pavlovic, "Punda: Probabilistic unsupervised domain adaptation for knowledge transfer across visual categories," in *ICCV*, 2017, pp. 3581–3590.

[196] A. Karbalayghareh, X. Qian, and E. Dougherty, "Optimal bayesian transfer learning," *IEEE Trans. Signal Processing*, vol. 66, no. 14, pp. 3724–3739, 2018.

[197] V. Perrone, R. Jenatton, M. Seeger, and C. Archambeau, "Scalable hyperparameter transfer learning," in *NIPS*, 2018.

[198] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf, "Large scale multiple kernel learning," *Journal of Machine Learning Research*, vol. 7, pp. 1531–1565, 2006.

[199] D. Xu and S. Chang, "Video event recognition using kernel methods with multilevel temporal alignment," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1985–1997, 2008.

[200] L. Duan, D. Xu, and I. Tsang, "Domain adaptation from multiple sources: A domain-dependent regularization approach," *IEEE Trans. Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 504–518, 2012.

[201] L. Dalton and E. Dougherty, "Optimal classifiers with minimum expected error within a bayesian framework-part i: Discrete and gaussian models," *Pattern Recognition*, vol. 46, no. 5, pp. 1288–1300, 2013.

[202] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks," in *NIPS*, 2014.

[203] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *ICML*, 2014.

[204] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *ICML*, 2015, pp. 97–105.

[205] M. Long, Y. Cao, Z. Cao, J. Wang, and M. Jordan, "Transferable representation learning with deep adaptation networks," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2018.

[206] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *arXiv*, 2015.

[207] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

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

[209] L. Liu, W. Lin, L. Wu, Y. Yu, and M. Yang, "Unsupervised deep domain adaptation for pedestrian detection," in *ECCV*, 2016.

[210] M. Long, H. Zhu, J. Wang, and M. Jordan, "Deep transfer learning with joint adaptation networks," in *ICML*, 2017.

[211] M. Long, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," *NIPS*, 2016.

[212] W. Ge and Y. Yu, "Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning," in *CVPR*, 2017, pp. 1086–1095.

[213] X. Zhang, F. Yu, S. Wang, and S. Chang, "Deep transfer network: Unsupervised domain adaptation," in *arXiv*, 2015.

[214] S. Motiian, M. Piccirilli, D. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *ICCV*, 2017, pp. 5715–5725.

[215] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *ICML*, 2011.

[216] M. Chen, Z. Xu, K. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," in *ICML*, 2012.

[217] F. Zhuang, X. Cheng, P. Luo, S. Pan, and Q. He, "Supervised representation learning: Transfer learning with deep autoencoders," in *IJCAI*, 2015, pp. 4119–4125.

[218] L. Wen, L. Gao, and X. Li, "A new deep transfer learning based on sparse auto-encoder for fault diagnosis," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 49, no. 1, pp. 136–144, 2019.

[219] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *ECCV*, 2016, pp. 597–613.

[220] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *NIPS*, 2004, pp. 529–536.

[221] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *ICML*, 2008.

[222] S. Kullback, "Letter to editor: the kullback-leibler distance," 1987.

[223] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *arXiv*, 2015.

[224] P. Pinheiro, "Unsupervised domain adaptation with similarity learning," in *CVPR*, 2018, pp. 8004–8013.

[225] J. Zhang, Z. Ding, W. Li, and P. Ogunbona, "Importance weighted adversarial nets for partial domain adaptation," in *CVPR*, 2018, pp. 8156–8164.

[226] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," in *AAAI*, 2018.

[227] W. Zhang, W. Ouyang, W. Li, and D. Xu, "Collaborative and adversarial networks for unsupervised domain adaptation," in *CVPR*, 2018, pp. 3801–3809.

[228] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster r-cnn for object detection in the wild," in *CVPR*, 2018, pp. 3339–3348.

[229] Q. Duan and L. Zhang, "Advnet: Adversarial contrastive residual net for 1 million kinship recognition," in *ACM MM Workshop on RFIW*, 2017, pp. 21–29.

[230] Q. Duan, L. Zhang, and W. Jia, "Adv-kin: An adversarial convolutional network for kinship verification," in *CCBR*, 2017, pp. 48–57.

[231] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, "Pairwise confusion for fine-grained visual classification," in *ECCV*, 2018, pp. 48–57.

[232] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," *ICCV*, vol. 30, no. 31, pp. 4068–4076, 2015.

[233] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand, "Domain-adversarial neural network," in *arXiv*, 2015.

[234] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," 2016.

[235] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *CVPR*, 2017, pp. 7167–7176.

[236] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto, "Few-shot adversarial domain adaptation," in *NIPS*, 2017.

[237] A. Rozantsev, M. Salzmann, and P. Fua, "Residual parameter transfer for deep domain adaptation," in *CVPR*, 2018, pp. 4119–4125.

[238] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *NIPS*, 2018.

[239] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *ICML*, 2018.

[240] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *CVPR*, 2018, pp. 3723–3732.

[241] J. Hoffman, E. Tzeng, T. Park, and J. Zhu, "Cycada: Cycle-consistent adversarial domain adaptation," in *arXiv*, 2017.

[242] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *CVPR*, 2017, pp. 3722–3731.

[243] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *ICLR*, 2017.

[244] L. Hu, M. Kan, S. Shan, and X. Chen, "Duplex generative adversarial network for unsupervised domain adaptation," in *CVPR*, 2018, pp. 1498–1507.

[245] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, "Image to image translation for domain adaptation," in *CVPR*, 2018, pp. 4500–4509.

[246] W. Hong, Z. Wang, M. Yang, and J. Yuan, "Conditional generative adversarial network for structured domain adaptation," in *CVPR*, 2018, pp. 1498–1507.

[247] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *CVPR*, 2018, pp. 79–88.

[248] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *CVPR*, 2018, pp. 5157–5166.

[249] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017.

[250] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," *California Institute of Technology*, 2007.

[251] C. Rate and C. Retrieval, "Columbia object image library (coil-20)," *Tech. Rep.*, 2011.

[252] D. Li, Y. Yang, Y. Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *ICCV*, 2017, pp. 5543–5551.

[253] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," pp. 2551–2559, 2015.

[254] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *CVPR*, 2009, pp. 951–958.

[255] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *CVPR*, 2017, pp. 5385–5394.

[256] B. Caputo, H. Mller, J. Martinez-Gomez, M. Villegas, B. Acar, N. Patricia, N. Marvasti, S. skdarl, R. Paredes, and M. Cazorla, *ImageCLEF 2014: Overview and Analysis of the Results*. Springer International Publishing, 2014.

Lei Zhang (M'14-SM'18) received his Ph.D degree in Circuits and Systems from the College of Communication Engineering, Chongqing University, Chongqing, China, in 2013. He was selected as a Hong Kong Scholar in China in 2013, and worked as a Post-Doctoral Fellow with The Hong Kong Polytechnic University, Hong Kong, from 2013 to 2015. He worked as a visiting professor at University of Macau in 2018. He is currently a Professor/Distinguished Research Fellow with Chongqing University. He has authored more than 70 scientific papers in top journals and conferences, including IEEE Transactions (e.g., T-NNLS, T-IP, T-MM, T-IM, T-SMCA), ICCV, ACM MM, AAAI, etc. His current research interests include machine learning, transfer learning, domain adaptation, computer vision and intelligent systems. Dr. Zhang was a recipient of the Best Paper Award of CCBR2017, the Outstanding Paper Award of Chongqing Association in Science and Technology, the Outstanding Reviewer Award of 9 international journals (e.g., pattern recognition, neurocomputing, etc.), Outstanding Doctoral Dissertation Award of Chongqing, China, in 2015, Hong Kong Scholar Award in 2014, Academy Award for Youth Innovation in 2013 and the New Academic Researcher Award for Doctoral Candidates from the Ministry of Education, China, in 2012.