Transfer Adaptation Learning: A Decade Survey
Lei Zhang, Senior Member, IEEE, and Xinbo Gao, 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 transfer adaptation learning (TAL) when it needs knowledge correspondence between different moments/domains. TAL 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 field of increasing influence and importance, which is presenting a blowout publication trend. This article surveys the advances of TAL methodologies in the past decade, and the technical challenges and essential problems of TAL have been observed and discussed with deep insights and new perspectives. Broader solutions of TAL being created by researchers are identified, i.e., instance reweighting adaptation, feature adaptation, classifier adaptation, deep network adaptation, and adversarial adaptation, which are beyond the early semisupervised and unsupervised split. The survey helps researchers rapidly but comprehensively understand and identify the research foundation, research status, theoretical limitations, future challenges, and understudied issues (universality, interpretability, and credibility) to be broken in the field toward generalizable representation in open-world scenarios.

Index Terms—Distribution discrepancy, domain adaptation (DA), generalizable representation, transfer learning (TL).

I. 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 based on a well-labeled image dataset, which can be recognized as target data (task). However, labeling a large number of target samples is cost-ineffective, because it consumes a lot of human resources and time expenses in labeling and becomes almost unrealistic for many domain-specific tasks in real applications. Therefore, leveraging another sufficiently labeled, distribution different but semantic-related source domain for recognizing target samples is becoming an increasingly important topic.

With the explosive increase in multisource 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 training a learning model using multisource 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, and weather condition in computer vision. Therefore, the classification performance on the target task is dramatically degraded when the source data for training the classifier have different distribution from the target data. This is because the fundamental independent identical distribution (i.i.d.) condition implied in statistical machine learning is no longer satisfied, which therefore promotes the emergence of transfer learning (TL) and domain adaptation (DA) [3]–[5]. Originally, the concept of transfer was first proposed in 1991 by Pratt et al. [6] and further introduced in [7] for transfer between neural networks. Mathematically, given the data and label pair (X, Y), the earlier TL supposed different joint probability distribution, i.e., P(Xsource, Ysource) ≠ P(Xtarget, Ytarget) between the source and target domains. DA supposed different marginal distribution, i.e., P(Xsource) ≠ P(Xtarget) but similar category space between domains, i.e., P(Ysource|Xsource) = P(Ytarget|Xtarget). Currently, a preliminary split of semisupervised domain adaptation (SSDA), unsupervised domain adaptation (UDA), open-set domain adaptation (OSDA), and partial domain adaptation (PDA) is presented, in all which the source labels are completely available. Notably, both SSDA and UDA assume the same category space C (close set), i.e., Csource = Ctarget. The former assumes a few labeled target samples are known, while the target labels are completely unknown for the latter. Contrastively, OSDA and PDA assume different category space across domains, i.e., Csource ≠ Ctarget. In general, OSDA supposes Csource ⊂ Ctarget and PDA supposes Csource ⊂ Ctarget.

The earlier related reviews on TL and DA can be referred to as [4], [8]–[17], which are summarized in Table I. However, our concern is that these existing surveys can only reflect partial work in TL/DA community. With the recent explosive increase in TL and DA models and algorithms, these surveys are unable to characterize, particularly the status, challenges, and future for this community from a more comprehensive perspective. The objective of this survey is much complete from concept, history, theory, models, in-depth insights, current focuses, application and benchmarks, and future outlook. Therefore, in this survey, we use a general name transfer adaptation learning (TAL) for unifying both TLs and DAs from a broader perspective and discuss the technical advances, challenges, and understudied issues. In the past decade, especially the recent five years, TAL was one of the most active areas in machine learning community, the
TABLE I
SUMMARIZATION OF RELATED SURVEYS IN TL/DA

| No. | Survey Title                                              | Ref. | Year |
|-----|----------------------------------------------------------|------|------|
| 1   | A Survey on Transfer Learning                           | [4]  | 2010 |
| 2   | Extreme learning machine based transfer learning algorithms: A survey | [8]  | 2017 |
| 3   | A survey of transfer learning                           | [9]  | 2016 |
| 4   | Transfer Learning for Visual Categorization: A Survey   | [10] | 2015 |
| 5   | A survey of transfer learning for collaborative recommendation with auxiliary data | [11] | 2016 |
| 6   | Visual Domain Adaptation                                 | [12] | 2015 |
| 7   | Transfer learning using computational intelligence: A survey | [13] | 2015 |
| 8   | Deep visual domain adaptation: A survey                 | [14] | 2015 |
| 9   | Transfer learning for activity recognition: a survey     | [15] | 2013 |
| 10  | Domain adaptation for visual applications: A comprehensive survey | [16] | 2017 |
| 11  | A Survey on Deep Transfer Learning                      | [17] | 2018 |

Fig. 1. Cross-domain object detection, recognition, and semantic segmentation. F denotes models of three tasks learned on the source domain. The value of TAL depends on the performance gain of target tasks.

goal of which is to narrow down the marginal and conditional distribution gap between the source and target data, such that the labeled source data from one or more relevant domains can be used for executing different target tasks. As shown in Fig. 1, the difference in visual cues across domains leads to a significant impact on visual understanding what the image conveys, owing to the classification vs. regression model bias caused by domain shift.

Moving forward, deep learning (DL) techniques [18]–[21] have recently become dominant and powerful algorithms in feature representation and abstraction for image classification. In particular, parameter adaptability and generality of DL models to other target data is worthy of praise, by fine-tuning a pretrained deep neural network (DNN) using a small amount of labeled target data. Therefore, fine-tune has become a commonly used strategy for training deep models and frameworks in various applications, such as object detection [22]–[25], person reidentification [26]–[28], medical imaging [29]–[31], and remote sensing [32]–[35]. Generally, fine-tune can be recognized as a prototype for bridging the sufficiently labeled big source data and the few/unlabeled target data [36], which also facilitates the research progress of TAL in computer vision. Conceptually, fine-tune is an intuitive and data-driven TL method, which depends on pretrained models with task-specific source database (e.g., ImageNet). The context of TL challenge and why the generated representation with pretrained models is useful have been formally explored in [36]. Extensively, from the viewpoint of generative learning, the popular generative adversarial net (GAN) [37] and its variants [38]–[42] that aim to synthesize plausible images belonging to target distribution from some source distribution (e.g., noise signal) can also be recognized as generalized TL techniques (e.g., style transfer tasks). Differently, conventional TL approaches put emphasis on output knowledge adaptation (e.g., high-level features) across the source and target domains, while GANs focus on input data adaptation (e.g., low-level image pixels) from source distribution to target distribution. Recently, image pixel-level transfer has been intensively studied in image-to-image translation [43]–[47], style transfer [48]–[50] and target face synthesis (e.g., pose transfer vs. age transfer) [51]–[55], etc.

In this survey, we focus on elaborating the technical advances and remained challenges in model-driven TAL. 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. Notably, TAL is similar but different from multitask learning that resorts to shared feature representations or classifiers for multiple different tasks [56], [57].

In the past decade, a number of TL and DA approaches have been emerged by following the expected target error upper bounds of DA that minimizes a convex combination of the source empirical risk, domain discrepancy, and joint error. This has been analyzed and justified [58], [59]. Specifically, given a labeled source domain $D_s$, an unlabeled target domain $D_t$ and a hypothesis $h \in \mathcal{H}$ (hypothesis set), according to the Ben–David theorem [58], the expected target risk $\epsilon_T(h)$ can be upper bounded by the expected source risk $\epsilon_S(h)$, domain discrepancy $d_{\mathcal{H}\mathcal{A}_T}(D_s, D_t)$ and joint error $\lambda = \min_{h \in \mathcal{H}} \epsilon_S(h) + \epsilon_T(h)$. Intuitively, DA can be achieved by modeling and simultaneously minimizing the upper bound of $\epsilon_T(h)$ from multiple different aspects, such as source error $\epsilon_S$, domain discrepancy $d(D_s, D_t)$, and joint error $\lambda$ under an ideal hypothesis $h^*$. Recently, Zhong et al. [60] proposed an E-MixNet to control the joint error. The open set and source-free DA theory are also presented in [61], [62]. Very recently, Fang et al. [63] build the generalization error theory of semisupervised heterogeneous DA and explain why labeled heterogeneous source data and unlabeled target data help reduce the target risk.

Due to endless emergence and variety of existing methods of visual transfer adaptation learning (VTAL), in this article, the remained challenges and main stream methodological advances in VTAL community are identified and surveyed without further discussing the DA theory. Specifically, around the ultimate challenge of universal representation, considering the covariate shift, domain discrepancy, and class discrimination, we propose five main stream key technical categories of TAL, which are beyond the early split of semisupervised and unsupervised (close set) and recent split of open set and partial set. The novel taxonomy of VTAL is summarized as follows.

1) Instance Reweighting Adaptation: Due to the 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 in a nonparametric manner. Probability and Shannon entropy-based reweighting has become a common technique in UDA, OSDA, and PDA problems. The adaptive
Fig. 2. Road map and research context of TAL. We present feature adaptation as the main stream consisting of three stages (1991~1993, 2006~2015, and 2015~2020). Other four subbranches, including instance reweighting, classifier adaptation, deep adaptation, and adversarial adaptation, are closely connected to the main stream toward universal representation. The milestones and key time nodes of TAL in the past decade have been flagged on each branch.

parameter estimation of the weights under a parametric distribution assumption remains to be a challenge.

2) Feature Adaptation: For adapting 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. However, learning common subspace generally belongs to shallow methodology, which relies more on pretrained deep models. It is challenging to unify the shallow model with the deep model. Also, the heterogeneity of data distribution makes it challenging to gain such generic feature.

3) Classifier Adaptation: The classifier trained on the source domain is often biased when recognizing instances from the target domain due to domain shift and category shift. Learning a generalized classifier for recognizing seen/partially seen/unseen object classes in another domain without bias is an essential but challenging topic.

4) Deep Network Adaptation: Large domain shift makes DNN difficult to obtain transferrable deep representation. The unavailability (uncertainty) of target labels (pseudo-labels) makes deep network adaptation challenging due to negative transfer.

5) Adversarial Adaptation: Adversarial learning is a general technique in TAL, originated from GANs. A basic objective of VTAL is to make the source and target domains more close in marginal distribution. Therefore, it is amounted to confusing the two domains, such that they cannot be easily discriminated. However, due to the class misalignment problem, there comes a technical challenge in class conditional domain confusion using adversarial training strategy.

This article provides a novel taxonomy of TAL regardless of closed set, open set, and partial set split. In the past 30 years, the overall research context of TAL, including the origin of transfer concept, establishment of DA theory, and rapid development of TAL models and algorithms toward universal representation, is illustrated in Fig. 2. The milestones with respect to those representative breakthroughs in theory and models around the taxonomy have been flagged.

II. INSTANCE REWEIGHTING ADAPTATION

When the training and test data are drawn from different distribution, this is commonly referred to as sample selection bias or covariate shift [64]. Instance reweighting aims to infer the resampling weight directly by feature distribution
matching across different domains in a nonparametric manner. Generally, given a dataset \((x, y) \sim P_t(x, y)\), a learning model can be obtained by minimizing the following expected risk of the training set:

\[
R(P_t, \theta, l(x, y, \theta)) = E_{(x,y)\sim P_t(\cdot, \cdot)}[l(x, y, \theta)].
\]

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

\[
R(P'_t, \theta, l(x, y, \theta)) = E_{(x,y)\sim P'_t(\cdot, \cdot)}[l(x, y, \theta)]
= E_{(x,y)\sim P'_t(x, y)}[\frac{P'_t(x, y)}{P_t(x, y)} l(x, y, \theta)]
= E_{(x,y)\sim P'_t(\cdot, \cdot)}[\beta(x, y) l(x, y, \theta)]
\]

where \(P_t(x, y)\) and \(P'_t(x, y)\) represent the probability distribution of the 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 \(P_t(x, y) = P'_t(x, y)\), we have \(\beta(x, y) = 1\).

From (2), we know that \(P_t(x, y)\) and \(P'_t(x, y)\) can be estimated for computing the weight \(\beta(x, y)\) by following [65] based on the prior knowledge of class distributions. Although this is intuitive, it requires very good density estimation of \(P_t(x, y)\) and \(P'_t(x, y)\). Particularly, a serious overweighting of observations with very large coefficients \(\beta(x, y)\) may be resulted from possible small errors or noise in estimating \(P_t(x, y)\) and \(P'_t(x, y)\). Therefore, to improve the reliability of weights, \(\beta(x, y)\) can be directly estimated by imposing flexible constraints into the learning model without having to estimate the two probability distributions.

The sample reweighting-based DA methods mainly focus on the case where the difference between the source domain and the target domain is not too large. The objective is to reweight the source samples so that source data distribution can be more close to target data distribution. Usually, when the distribution difference between the two domains is relatively large, the sample reweighting methods can be combined with others (e.g., feature adaptation) for auxiliary TL. Instance reweighting has been studied with different models, which can be divided into three categories based on weighting scheme: 1) intuitive weighting; 2) kernel-mapping-based weighting; and 3) co-training-based weighting. This kind of methods put emphasis on the learning or computation of the weights using different criterions and training protocols. The taxonomy of instance reweighting-based models is summarized in Table II.

### A. Intuitive Weighting

Instance reweighting-based DA was first proposed for natural language processing (NLP) [66], [67]. In [66], Jiang and Zhai proposed an intuitive instance weighted DA framework, which introduced four parameters for characterizing the distribution difference between the source and target samples. For example, for each \((x^t_i, y^t_i) \in D_t\), the labeled source data, the parameter \(a_t\) that was used to indicate how likely \(P_{target}(y^t_i| x^t_i)\) is close to \(P_{source}(y^t_i| x^t_i)\) and the parameter \(b_t\) that was ideally computed as \(P_{target}(x^t_i)/P_{source}(x^t_i)\) were introduced. Obviously, large \(a_t\) means high confidence of the labeled source sample \((x^t_i, y^t_i)\) contributing positively to learning effectiveness. Small \(a_t\) means the two probabilities are very different, and the instance \((x^t_i, y^t_i)\) can be discarded in the learning process. In addition, for each \(x^t_i, y^t_i \in D_t\), the unlabeled target data, and for each possible label \(y \in \mathcal{Y}\), the hypothesis space, the parameter \(y_i(y)\) that indicates how likely a tentative pseudo-label \(y\) can be assigned to \(x^t_i, y^t_i\), and then \((x^m_i, y)\) is included as a training sample.

Generally, \(a_t\) and \(y_i\) play an intuitive role in sample selection by removing those misleading source samples and adding those valuable labeled target samples during the TL process. Although 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 [66]. Therefore, adaptive learning of these intuitive weights still remains a challenging issue.

In [67], Wang et al. proposed two instance weighting schemes for neural machine translation (NMT) DA, i.e., sentence weighting and dynamic domain weighting. Specifically, given the parallel training corpus \(D = [D_{tn}, D_s]\) consisting of in-domain data and out-of-domain data, the sentence weighted NMT is written as

\[
J_{sw} = \sum_{(x, y) \in D} \lambda_i \log P(y|x).
\]

where \(\lambda_i\) is the weight to score each \((x_i, y_i)\). \(P(\cdot)\) 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 \(\lambda\) was designed for the in-domain data, and the NMT objective function (3) can be transformed as [67]

\[
J_{dw} = \lambda \sum_{(x, y) \in D_m} \log P(y|x) + \sum_{(x', y') \in D_o} \log P(y'|x').
\]

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 [68] TL method, which leveraged the boosting algorithm to automatically tune the weights of the training samples.

In [69], Chen et al. proposed a more intuitive weighting-based subspace alignment (SA) method by reweighting the source samples for generating source subspace that are close to the target subspace. Let \(w = [w_1, \ldots, w_m]^T \in \mathcal{R}^m\) denote the weighting vector of the source samples. Obviously, \(w_i\) with respect to the source sample \(x_i\) increases if its distribution is more close to the target data. Therefore, a simple weight assignment strategy was presented for assigning larger weights to the source samples that are closer to the target domain [69].

After obtaining the weight vector \(w\), the weighted source space can be obtained by performing principal component
analysis (PCA) on the following covariance matrix $C$ of weighted source data:

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

(5)

where $\mu$ 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 UDA model, SA [76], with Frobenius norm minimization, was implemented:

$$
\min_{M} \| P_{SM} - P_{T} \|_F^2
$$

(6)

where $M$ can be easily solved with least-square solution.

B. Kernel-Mapping-Based Weighting

The intuitive weighting-based DA was implemented in the raw data space. To infer the sampling weights by distribution matching across the source and target data in feature space in a nonparametric way, kernel-mapping-based weighting was proposed. Briefly, the distribution difference can be better characterized by reweighting the source samples such that the means of the source and target instances in a reproducing kernel Hilbert space (RKHS) are close [64]. Kernel-mapping-based weighting consists of two categories of methods: distribution matching [64], [70], [71] and sample selection [72], [73].

1) Distribution Matching. The intuitive idea of distribution matching is to match the means between the source and target data in a RKHS by resampling the weights of the source data. Two similar distribution matching criteria, i.e., kernel mean matching (KMM) [64] and maximum mean discrepancy (MMD) [77], [78], have been used as nonparametric statistic to measure the distribution difference. Specifically, Huang et al. [64] first proposed to reweight the source samples with $\beta$, such that KMM between the means of target data and the weighted source data is minimized

$$
\min_{\beta} \| E_{x \sim P_S} [\phi(x)] - E_{x \sim P_T} [\beta(x)\phi(x)] \|_F
$$

s.t. $\beta(x) \geq 0, E_{x \sim P_T} [\beta(x)] = 1$

(7)

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

Chu et al. [70] 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

$$
(w, s) = \arg \min_{w, s} \{ R_w(D^w, s) + \lambda \Omega_s(D^w, D^{rec}) \}
$$

(8)

where $R_w(\cdot)$ is the empirical risk (loss) on the training set $D^w$, $\Omega_s(\cdot)$ indicates the distribution mismatch formulated by KMM, $s$ is the weighting vector of the source samples, and $w$ is the classifier parameters. From (8), the KMM-based distribution mismatch plays an important role in model regularization on the sampling weights.

More recently, Yan et al. [71] proposed a weighted MMD (WMMD) for DA, 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 the source and target domains. WMMD is formulated as [71]

$$
d^2_{wmmd} = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N} \alpha_y \phi(x_i^y) - \frac{1}{N} \sum_{j=1}^{N} \phi(x_j^y))^2
$$

(9)

where $\alpha_y$ is the class weight with respect to class $y$ of the $i$th source sample and $\phi(\cdot)$ is the nonlinear mapping into RKHS $\mathcal{H}$. $M$ and $N$ denote the number of samples drawn from the source and target domains, respectively.

2) Sample Selection is another kind of kernel-mapping-based reweighting method. Zhong et al. [72] proposed a cluster-based sample selection method KMapWeighted which was established on the assumption that 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 the target data via a $K$-means-based clustering criterion. The data in the same cluster should be with the same labels and then target-like source samples were selected.

Long et al. [73] proposed a transfer joint matching (TJM) method for DA by minimizing the MMD-based distribution mismatch between the 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 strong correlation between the source samples and the target domain samples. The TJM model is provided as [73]

$$
\min_{A^{TMA}=I} \tr(A^{TMA}) + \lambda \left( \| A \|_{2,1} + \| A \|_F^2 \right)
$$

(10)

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

C. Co-Training-Based Weighting

Co-training [79] 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 TL, Chen et al. proposed a CODA [74] method, in which two classifiers with different weight vectors were trained. For better training both the classifiers on the training set, the two classifiers were jointly minimized with weighting. In essence, the method of sample reweighting based on the classifier is similar to TrAdaBoost [68] and KMapWeighted [72].

In [75], Chen et al. proposed a reweighted adversarial adaptation network (RAAN) for UDA. Two classifiers including a multiclass source instance classifier $C$ and a binary domain classifier $D$ were designed for adversarial training. The domain classifier $D$ aims to discriminate whether features are from the source or target domain, while the domain feature representation network $T$ tries to confuse them, which formulates an adversarial training manner. For improving the domain confusion effect, the source feature distribution is reweighted with $\beta$ during training of the domain classifier $D$. With the gaming between $T$ and $D$ as GAN does [37], the minimax objective is formulated [75]

$$
\min_{T} \max_{D, \beta} \mathcal{L}_{adv}
$$

(11)
TABLE III
OUR TAXONOMY OF FEATURE ADAPTATION APPROACHES

| Feature Adaptation | Model Basis     | Reference         |
|-------------------|----------------|-------------------|
| Geodesic path     | Grassman manifold | [80], [81]       |
| Alignment         | Subspace learning | [76], [82],       |
|                   |                  | [83], [84]       |
| Feature Transformation |                |                   |
| Projection        | MMD&Hsic&       | [85], [86],       |
|                   | Bregman divergence | [87], [88]     |
| Metric            | First/second-order statistic | [89], [90], |
|                   |                  | [91], [92]       |
| Augmentation      | Zero-padding&    | [93], [94],       |
|                   | Generative      | [95], [96]       |
| Feature Reconstruction |                |                   |
| Low-rank models   | Low-rank representation (LRR) | [97], [98], |
|                   |                  | [96], [99]       |
| Sparse models     | Sparse subspace clustering (SSC) | [100], [101], [102], [103] |
| Feature Coding    |                  |                   |
| Domain-shared dictionary | Dictionary learning | [104], [105], |
|                   |                  | [106], [107]     |
| Domain-specific dictionary | Dictionary learning | [108], [109], |
|                   |                  | [110], [111]     |

where the weight $\beta$ is multiplied with $D$ and trained in a cooperative way. The learning of the source classifier $C$ was easily performed by minimizing the cross-entropy loss.

D. Discussion and Summary

We recognize three kinds of instance reweighting: intuitive, kernel mapping, and co-training. Although instance reweighting 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 and 2) the initialization and estimation of instance weights are important and can be treated as a latent variable obeying some distribution.

III. FEATURE ADAPTATION

Feature adaptation aims to discover common feature representation of the data drawn from multiple sources using different techniques including linear and nonlinear ones. In the past decade, feature-adaptation-induced TAL has been intensively studied, which, in our taxonomy, can be categorized into 1) feature-subspace-based; 2) feature-transformation-based; 3) feature-reconstruction-based; and 4) feature-coding-based. Despite these advances, the technical challenges being faced by researchers lie in the domain SA, projection learning for distribution matching, generic representation, and shared domain dictionary coding. The taxonomy of feature adaptation approaches is summarized in Table III.

A. Feature-Subspace-Based

Learning subspace generally resorts to UDA. Three representative models are referred to as sampling geodesic flow (SGF) [80], geodesic flow kernel (GFK) [81], and SA [76]. There exists a common property of the three methods, i.e., the data are assumed to be represented by a low-dimensional linear subspace. That is, a low-dimensional Grassmann manifold is embedded in the high-dimensional space. Generally, 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. [80] is an unsupervised low-dimensional subspace transfer method, which samples a group of subspaces along the geodesic path between domains, and aims to find an intermediate representation with closer domain distance.

Similar to but different from SGF, Gong et al. proposed a GFK [81], in which GFK 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 semidefinite GFK as follows:

$$\langle z_i^\infty, z_j^\infty \rangle = \int_0^1 (\phi(t)^T x_i)^T (\phi(t)^T x_j) dt = x_i^T G x_j$$

(12)

where $G$ is a positive semidefinite mapping matrix. With $z = (G)^{1/2} x$, features in the original space can be transformed into the Grassmann manifold space.

For aligning the source subspace to the target subspace, in SA [76], 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 (6). As presented in SA, the subspaces of the source and target data were spanned by the eigenvectors induced with a PCA. Furthermore, Sun and Saenko proposed a subspace distribution alignment (SDA) [82] by simultaneously aligning the distributions and 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) [83] framework, which introduced a more generic method for moving the source subspace $W_i$ closer to target subspace $W_j$.

$$\min_{W_i, W_j} \frac{1}{2} \left\| M^{1/2} \odot (W_i - W_j) \right\|^2$$

(13)

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. Furthermore, a guide subspace learning (GSL) was proposed for UDA [84], and an online continual DA with evolving domains was proposed in [112].

B. Feature-Transformation-Based

This kind of models aim to learn a transformation or projection of the data with some distribution matching metrics between the source and target domains [5], [113], [114]. Then, the transformed or projected feature distribution difference across two domains can be removed or relieved. Feature-transformation-based DA has been a main stream in visual TL community in the past years, which can be further divided.
into projection, metric, and augmentation according to model formulation.

1) Projection-Based DA aims to solve a projection matrix in the source and target domains for reducing the marginal distribution difference and conditional distribution difference between domains, by introducing kernel matching criterion [85]–[88], [115]–[118] and discriminative criterion [119]–[122]. The kernel matching criterion generally adopts the MMD statistic, which characterizes the marginal distribution difference and conditional distribution difference between the source and target data. In UDA setting, the labels of target domain samples are generally unavailable, and therefore the pseudo-labels of target samples should be iteratively predicted for quantifying the conditional MMD between domains [119], [123]. The discriminative criterion focuses on within-class compactness and between-class separability of the projection. Mathematically, the formulation of empirical nonparametric MMD in universal RKHS is written as

\[ d^2_{\gamma}(D_1, D_2) = \frac{1}{M} \sum_{i=1}^{M} \phi(x^i_1) - \frac{1}{N} \sum_{j=1}^{N} \phi(x^j_2) \]  

(14)

Specifically, with MMD-based kernel matching criterion, Pan and Yang first proposed a transfer component analysis (TCA) [85] by introducing the marginal MMD as projection loss function. The joint distribution adaptation (JDA) proposed by Long et al. [86] 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

\[ \min_W d^2_{\text{MMD}}(X_S, X_T, W) + \lambda d^2_{\text{MMD}}(Y_S, Y_T, W) \]  

(15)

where \( W \) denotes the projection matrix, \( Y_T \) denotes the predicted pseudo-label of the target data, and \( d^2_{\text{MMD}} \) 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 [119] and CDSL [120], and generally written as

\[ \min_W F(W, X_S, X_T, Y_S, Y_T) + \lambda d^2_{\text{MMD}}(X_S, X_T, W) \]  

(16)

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), and PCA. In addition to the MMD criterion for projection-based transfer models, Bregman divergence [88], Hilbert–Schmidt independence criterion (HSIC) [87], [103], [124], and manifold criterion [96] have also been presented.

In [88], 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

\[ W = \arg \min_W F(W) + \lambda D_W(P_L||P_U) \]  

(17)

where \( F(W) \) is similar to (16) 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. [124], the same author as that of MMD, was used to measure the dependency between two sets \( \mathcal{X} \) and \( \mathcal{Y} \). Let \( k_s \) and \( k_t \) denote the kernel function w.r.t. the RKHS \( \mathcal{F} \) and \( \mathcal{G} \). The HSIC is mathematically written as [124]

\[ \text{HSIC}(\mathcal{X}, \mathcal{Y}, \mathcal{F}, \mathcal{G}) = \| C_{\mathcal{X} \mathcal{Y}} \|_{F}^{2} = (N - 1)^{-2} \text{Tr}(K_{\mathcal{X}} H K_{\mathcal{Y}} H) \]  

(18)

where \( N \) is the size of the set \( \mathcal{X} \) and \( \| C_{\mathcal{X} \mathcal{Y}} \|_{F}^{2} \) is the Hilbert–Schmidt norm of the cross-covariance operator. \( K_{\mathcal{X}} \) and \( K_{\mathcal{Y}} \) denote the two kernel Gram matrices, and \( H \) is the centering matrix. HSIC will be zero if and only if \( \mathcal{X} \) and \( \mathcal{Y} \) are independent. In [103], Wang et al. proposed to use the projected HSIC as regularization, which is written as

\[ \min_W F(W) - \lambda \text{HSIC}(\mathcal{X}, \mathcal{Y} = \mathcal{F}, \mathcal{G}) \]  

(19)

where \( \mathcal{Y} \) denotes the label set of the 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 the 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 [83], [92], [99], [119]. For moving the two projections of both the domains closer, the Frobenius norm of their difference like (13) 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 [126]. Metric transfer has a close link to projection-based examples, if the metric \( M \) is a semidefinite matrix and can be decomposed into \( M = W W^T \) [89]. The metric-based transfer can be divided into first-order statistic [89], [90], [113], [127]–[130] and second-order statistic [91], [92], [131]–[133]-based distance metric, such as Euclidean, Mahalanobis distance, and moment distance.

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

\[ \min_{M} d(M, \phi(X_S), \phi(X_T)) + \lambda \| M \| \]  

(20)

where \( \phi(\cdot) \) is the feature representation or mapping function, and it can be linear mapping [113], kernel mapping [90], [128], autoencoder [89], or neural network [127].

For example, robust transfer metric learning (RTML) proposed by Ding and Fu [89] adopted an autoencoder-based feature representation for metric learning, such that the Mahalanobis distance between the source and target domains is minimized. The objective function of RTML is as follows:

\[ \min_{M \in S^2} \sum_{i=0}^{c} \text{tr}(\phi_i M) + a \| X - M \tilde{X} \|_{F}^2 + \lambda \text{rank}(M) \]  

(21)

where \( M \) is a positive semidefinite matrix, \( \tilde{X} \) is the repeated version of \( X \), and \( \tilde{X} \) is the randomly corrupted version of \( \tilde{X} \). The first item is the Mahalanobis-distance-induced domain discrepancy under metric \( M \), the second item is the autoencoder, 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 the
source and target domains instead of the means is minimized [91], [92], [131], [132]. For example, Sun et al. [91], [131] proposed a simple but efficient correlation alignment (CORAL) by aligning the second-order statistic (i.e., the covariance) between the source and target distributions instead of the first-order metric. By introducing a metric matrix $A$, the difference between source covariance $\Sigma_S$ and target covariance $\Sigma_T$ in CORAL can be minimized by solving

$$\min_A \| A^T \Sigma_S A - \Sigma_T \|_F^2. \quad (22)$$

Equation (22) is amount to matching the two centered Gaussian distribution, which is the basic assumption for such second-order statistic-based transfer.

3) Augmentation-Based DA often assumes 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 generic zero-paddling [93], [94], [134]–[136] and the latest generative [95], [96], [137] types.

Zero-paddling was first proposed by Daume III [134], which presented an Easy Adapt (EA) model. Assume the raw input data space to be $\mathcal{X} \in \mathbb{R}^F$, then the augmented feature spaces should be $\mathcal{Y} \in \mathbb{R}^M$. By defining the mapping functions of the source and target domains from $\mathcal{X}$ to $\mathcal{Y}$ as $\Phi_s(\cdot)$ and $\Phi_t(\cdot)$, respectively. Then is

$$\Phi_s(x) = [x, x, 0], \quad \Phi_t(x) = [x, 0, x] \quad (23)$$

where $0 \in \mathbb{R}^F$ is a zero vector. The first, second, and third bits of the augmented feature $\Phi(x)$ in (23) represent the common, source-specific, and target-specific feature components, respectively. However, in heterogeneous DA that addressing different feature dimensions between the source and target domains [138]–[140], for example, cross-modal learning (e.g., images vs. text), Li et al. [94] argued that such simple zero-paddling for dimensionality consistence between domains is not meaningful. The reason is that there would be no correspondences between heterogeneous features. Therefore, Li et al. [94] 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 $P \in \mathbb{R}^{d_s \times d_i}$ and $Q \in \mathbb{R}^{d_t \times d_i}$. The augmented feature for the source and target domains can be written as

$$\Phi_s(x_s) = [P x_s, x_s, 0_d], \quad \Phi_t(x_t) = [Q x_t, 0_d, x_t] \quad (24)$$

where $d_i$ and $d_r$ represent the dimensionality of the source and target data, respectively. For incorporating the unlabeled target data, Daume III further proposed an EA++ model with zero-paddling-based feature augmentation for semisupervised DA [93], [135]. Chen et al. [136] proposed to use zero-paddling-based camera correlation-aware feature augmentation (CRAFT) for cross-view person reidentification.

Generative methods used for feature augmentation mainly focus on plausible data generation toward enhancing the robustness of domain transfer. In [95], Volpi et al. proposed an adversarial feature augmentation by introducing two 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 the 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 the domains. In [96], 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.

C. Feature-Reconstruction-Based

Feature reconstruction 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 the source and target domains, while excluding noises and outliers during DA. To this end, feature-reconstruction-based domain transfer can be generally divided into three types: low-rank reconstruction [96]–[99], sparse reconstruction [100]–[103], and graph matching [141]–[146]. For the first one, 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 second one, sparsity or structural sparsity was generally used for transferrable sample selection. For the last one, the graphs of each domain are matched by a correspondence matrix. Methodologically, reconstruction-based domain transfer is closely related to low-rank representation (LRR) [147], [148], matrix recovery [149], [150], and sparse subspace clustering (SSC) [151]–[153].

1) Low-rank reconstruction-based DA was first proposed by Jhuo et al. [97], in which $W$ 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

$$\min_{W,Z,E} \text{rank}(Z) + \alpha \| E \|_{2,1}$$

s.t. $WX_S = X_T Z + E, \quad WW^T = I. \quad (25)$

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 [97], Shao et al. [98] proposed a latent subspace transfer learning (LTSL), which tends to reconstruct the target data using the source data as basis in a projected latent subspace

$$\min_{W,Z,E} F(W, X_S) + \lambda_1 \text{rank}(Z) + \alpha \| E \|_{2,1}$$

s.t. $W^T X_S = X_T Z + E \quad (26)$

where $F(\cdot)$ is a subspace learning function, similar to (16), (17) and (19). By comparing (25) with (26), the major difference lies in latent space learning of $W$ for both the domains under LTSL. Both the methods, established on LRR, advocated low-rank reconstruction between domains for TL. As demonstrated in [148], 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 SSC, which, different from LRR, is well-supported by theoretical analysis and experiments when handling data
near the intersections of subspaces [152]. Therefore, in [100], Zhang et al. proposed a latent sparse domain transfer (LSDT) model, which jointly learns the sparse coding $Z$ between domains and the latent subspace $W$

$$\min_{Z,W} \|Z\|_1 + \lambda_1 \|WX_T - WXZ\|_F^2$$

$$+ \lambda_2 \|X - WTWX\|_F^2$$

s.t. $WW^T = I$, $1_{N_T} + N_SZ = 1_{NY}^T$

$$Z_{N_T+i,i} = 0 \ \forall i = 1, \ldots, N_T$$

(27)

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 DA, which is more robust to noise or outliers drawn from the source domain. The model has also been kernelized by defining the projection $W$ as the linear representation of $X$. The reconstruction is then implemented in a high-dimensional RKHS, based on the representer theorem. In [102], Zhang et al. proposed a $l_2,1$-norm constraint-based reconstruction transfer model with discriminative subspace learning and domain-class consistency was guaranteed. The joint constraint with low-rankness and sparsity for the reconstruction matrix was proposed in [101], such that the global and local structures of data can be preserved.

3) Graph matching-based DA focused on the graph constructions of both the domains with adjacency matrices and solved the correspondence matrix $C$ between graphs for inherent domain similarity matching. In [141], the first-order (data matrices) and second-order (adjacency matrices)-based matching were formulated

$$\min_{C} \|CX_T - X_S\|_F^2 + \lambda\|CD_T - D_SC\|_F^2 + \gamma R(C)$$

(28)

where $D_S$ and $D_T$ represent the adjacency matrices of graphs for the source and target domains, respectively. $R(C)$ denotes the class regularization.

D. Feature-Coding-Based

In feature-reconstruction-based transfer models, the focus is learning of reconstruction coefficients across domains, on the basis of raw feature of the source or target data. Different from that, feature-coding-based TL 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 dataset under a sparsity constraint [154]–[156]. 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 [104]–[107] and domain-specific dictionary-based [108]–[111], [157]. Obviously, the former resorts to learning one common dictionary for both the 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 domains using a common dictionary. In [104], [106], 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

$$\min_{D,P,\alpha} \sum_{k \in \{s,t\}} \|P(k)X(k) - DA(k)\|_F^2 + \eta(D, P, \alpha)$$

(29)

where $P$ denotes the latent subspace projection, $\alpha$ denotes the representational coefficients for the source data $X_S$ and target data $X_T$ using a shared dictionary $D$, and $\eta(\cdot)$ denotes the regularizer. The shared dictionary $D$ is demonstrated to incorporate the common information from both the 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 [110], [157]. The general model can be written as

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

(30)

where $\Omega(\cdot)$ denotes the difference between representation coefficients of the source and the target. If $\alpha_s = \alpha_t = \alpha$, then $\Omega(\alpha_s, \alpha_t) = 0$ and the model in (30) is degenerated as the common representation-coefficient-based domain adaptive dictionary learning [108].

In [105], [109], [111], a set of intermediate domains that bridge the gap between domains were incorporated as multiple dictionaries $\{D_k\}_{k=0}^{K}$, which can progressively capture the intrinsic domain shift between source domain dictionary $D_0$ and target domain dictionary $D_K$. The difference $\Delta D_k$ between the atoms of adjacent two subdictionaries can well characterize the incremental transition and shift. Actually, this kind of models can be linked with SGF [80] and GFK [81] by sampling finite or infinite number of intermediate subspaces on the Grassmann manifold for better capturing the intrinsic domain shift.

E. Discussion and Summary

In this section, feature adaptation methods are presented, including subspace, transformation, reconstruction, and coding-based types. Feature adaptation has been intensively studied by addressing negative transfer and underadaptation problems from different perspectives. Two future directions in feature level are specified: 1) more reliable probability distribution similarity metric is needed, except the Gaussian kernel-induced MMD; and 2) for learning domain-invariant representation, model ensemble of linear and nonlinear ones is desired.

IV. CLASSIFIER ADAPTATION

In cross-domain visual categorization, classifier-adaptation-based TAL aims to learn a generic classifier by leveraging labeled samples drawn from the source domain and few labeled samples from the target domain [3], [158]–[160]. Typical cross-domain classifier adaptation can be divided into 1) kernel classifier-based [3], [158], [161]–[165], 2) manifold regularization-based [166]–[172], and 3) Bayesian classifier-based [173]–[178]. The taxonomy of classifier adaptation approaches is summarized in Table IV.

A. Kernel-Classifier-Based

Yang et al. [3] first proposed an adaptive support vector machine (ASVM) in 2007 for target classifier training, which assumed that there exists a bias $\Delta f(x)$ between the source classifier $f^a(x)$ and the target classifier $f(x)$. There is

$$f(x) = f^a(x) + \Delta f(x) = f^a(x) + w^T \phi(x)$$

(31)
where $w$ is the parameter of the bias function $\Delta f(x)$, which was solved by standard SVM

$$
\min_{w} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \varepsilon_i \\
\text{s.t. } y_i f^a(x_i) + y_i w^T \phi(x_i) \geq 1 - \varepsilon_i, \quad \varepsilon_i > 0.
$$

In (32), $f^a(\cdot)$ 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 [161]–[164], 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 [179], [180]. In addition, for reducing distribution mismatch, MMD-based kernel matching metric $\mathcal{D}_k^2(\cdot)$ was jointly minimized with the structural-risk-based classifiers. The general model of MKL-based classifier adaptation can be written as

$$
\min_{k,f} R(k, f, X_S, X_T) + \lambda \Omega(d_k^2(X_S, X_T))
$$

where $R(\cdot)$ denotes the structural risk on labeled training samples, $f$ is the decision function, $\Omega(\cdot)$ is the monotonic increasing function, and $l = \sum_{m=1}^{M} d_m k_m$ is a linear combination of a set of base kernels $k_m$ with $\sum_{m=1}^{M} d_m = 1$ and $d_m \geq 0$. The structural risk $R(\cdot)$ was generally formulated based on the hinge loss, i.e., $l_1(t) = \max(0, 1 - t)$, as that in SVM. Duan et al. [181] 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 [166], [167].

### B. Manifold-Regularizer-Based

The manifold assumption in semisupervised learning means that similar samples with small distance in feature space more likely belong 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. [166] and Cao et al. [169] 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 [170], 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 [167] proposed a manifold regularization-based least-square classifier EDA on both the domains with label precomputation and refining for DA. More recently, Wang et al. [168] 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

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

where $X$ is the data of the source and target domains, $F$ 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 $D_{ii} = \sum_j W_{ij}$. 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 [96].

### C. 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 UDA, an underlying assumption in the kernel classifier and manifold-classifier-based models is that the conditional domain shift can be minimized without relying on the target labels. In addition, 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 models, i.e., Bayesian classifier-based graphical models for DA/TL, have been studied [173]–[178], which aim to have better insights on the transfer process.

In [173], Gönen and Margolin first proposed graphical model, i.e., kernelized Bayesian transfer learning (KBTL) for DA. 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 [176], Gholami et al. proposed a probabilistic latent variable model (PUnDA) for UDA, 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 computation of exact posterior distribution of the latent variables is intractable. More recently, Karbalayghareh et al. [177] proposed an optimal Bayesian transfer learning (OBTL) classifier to formulate the optimal Bayesian classifier (OBC) in the target domain using the prior knowledge of the source and target domains, where OBC [182] aims to achieve Bayesian minimum mean squared error over uncertainty classes. To avoid costly computations (MCMC sampling), OBTL classifier was derived based on the Laplace approximated hypergeometric functions.

### D. Discussion and Summary

In this section, classifier adaptation including kernel classifier, manifold regularizers and Bayesian classifier are surveyed, which mostly rely on a small amount of tagged target domain

| Classifier Adaptation | Model Basis | Reference |
|-----------------------|-------------|-----------|
| Kernel Classifier     | SVM&MKL     | [3], [162], [163], [158], [164], [165] |
| Manifold Regularizer  | Label Propagation & MMD | [166], [167], [168], [169], [170], [172] |
| Bayesian Classifier   | Probabilistic graph models | [173], [174], [175], [176], [177], [178] |
data and facilitate semisupervised TL. This can be easily adapted to unsupervised TL by precomputing and iteratively updating the pseudo-labels of the completely unlabeled target domain in classifier adaptation. 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 negative effect; 3) the manifold assumption does not hold due to serious domain disparity.

V. DEEP NETWORK ADAPTATION

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. Yosinski et al. [183] have 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 [184], Donahue et al. proposed the deep convolutional activation feature (DeCAF) extracted using a pretrained AlexNet model [20], which has well-proved the generalization of DNNs for generic visual classification. This work further facilitated deep TL and deep DA. Generally, the presented three types of TAL models, including instance reweighting, feature adaptation, and classifier adaptation, can be incorporated into DNNs with end-to-end training for deep network adaptation. In 2015, Long et al. [185], [186] proposed a deep adaptation network (DAN) for learning transferable 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 higher layered features into RKHSs for nonparametric kernel matching (e.g., MMD-based) between domains. In the training process, DAN was trained by fine-tuning on ImageNet pretrained DNN, such as AlexNet [20], VGGNet [187], GoogLeNet [188], and ResNet [21].

Another subbranch of deep network adaptation is PDA [189]–[192] and OSDA [193]–[195], which loose the category space (label set) consistency assumption across domains. The former supposes that the source domain shares partial categories with the target data, i.e., $C_S \supset C_T$, and the disjoint source categories are recognized as noisy instances. The latter assumes that the target domain shares partial categories with source data, i.e., $C_S \subset C_T$, and the disjoint target categories are unified as unknown category, i.e., the $(C + 1)$th class. $C$ is the number of known classes. The common strategy is to actively discover the common/private classes with thresholding and reweighting the model at sample level. More general, You et al. [196] proposed a universal domain adaptation (UDA) which requires no prior knowledge on the label sets. That is, both the domains can have private categories. Actually, the methodology in conventional DA can be freely generalized into PDA or OSDA with modifications, and therefore, they are not further specified due to space limitation.

Various deep network adaptation models have emerged by modifying the network architecture, loss, and threshold-old tricks, which is presenting a blowout trend, because deep network adaptation yields significant performance gains against shallow DA methods. For a better overview of various deep network adaptation methods, we divide them into 1) marginal-alignment-based, 2) conditional-alignment-based, 3) batch normalization (BN) layer-derived, and 4) autoencoder-based, in which the first three mainly focus on CNN architecture. The taxonomy of deep network adaptation approaches is summarized in Table V.

### A. Marginal-Alignment-Based

In unsupervised deep DA frameworks, for reducing the distribution disparity $d_{H\times H}(D_S, D_T)$ between labeled source domain and unlabeled target domain, the top layered features were generally transformed to an RKHS space where MMD-based kernel matching between domains was performed, which is recognized as marginal-alignment-based deep network adaptation [185], [197]–[199]. 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. [197] and DAN [185]. The model can be written as

$$
\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} J(\theta(x_i), y_i) + \lambda \sum_{l} d_{ma}(D_l^s, D_l^t)
$$

where $J(\cdot)$ is the cross-entropy loss function, $\theta(\cdot)$ is the feature representation function, $D_l$ denotes the domain feature set from the $l$th layer, and $d_{ma}(\cdot)$ is the marginal alignment function [e.g., MMD in (14)] between domains. Clearly, in (35), 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. [199] in a tensor product Hilbert space for matching the joint distribution of activations of multiple layers. Besides, other domain discrepancy metrics, such as Kullback–Leibler (KL) divergence [219]–[222], Jensen–Shannon (JS) divergence [223], Wasserstein distance [224], [225], larger feature norms [226], margin disparity discrepancy [227], moment distance [228], and optimal transport distance [229], [230], have also been explored. Interestingly, a new but practical DA setting is proposed with a concept of complementary label in the source data [231], [232].

The model in (35) does not take into account the network outputs of target domain stream, which may not well adapt the source classifier to the target data. For addressing this problem, conditional entropy minimization principle [233] that favors the low-density separation between classes in unlabeled target data $D_t$ was further exploited in [186], [200], [201].
The entropy minimization is written as

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

(36)

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. In addition, by following the assumption of ASVM in [3], the residual $\Delta f(x)$ between the source and target classifiers was learned in the residual transfer network (RTN) [200], with a residual connection.

### B. Conditional-Alignment-Based

In marginal-alignment-based deep network adaptation, only the top layered feature matching in RKSH spaces was formulated using the nonparametric MMD metric. However, high-level semantic information was not taken into account in domain alignment, which may degrade the adaptability of level semantic information was not taken into account in the top layered feature matching in RKSH spaces was formulated using the nonparametric MMD metric. However, 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 due to class misalignment. Theoretically, an unexpected larger upper bound of the joint error $\lambda = \min_{h \in \mathcal{T}} (\xi(h, l_s) + \xi_T(h, l_t) + \varepsilon_T(l_s, l_t)) \leq \min_{h \in \mathcal{T}} \xi_S(h, l_s) + \xi_T(h, l_t) + \varepsilon_T(l_s, l_t)$ may be encountered, where $l_s$ and $l_t$ are the labeling function for the source and target domains, respectively. This is because $\varepsilon_T(l_s, l_t)$ can be large enough due to class misalignment. Therefore, conditional-alignment-based deep network adaptation methods were presented jointly with the marginal-alignment-based models [202]–[208]. Due to the unavailability of target labels, in such kinds of models, clustering labels or progressively updated target pseudo-labels (label refinement) are generally exploited. Similar to the formulation of MMD in (14), the conditional alignment was generally formulated by building between-domain MMD like discrepancy metric $d_{\mathcal{M}}^2$ on the probability $p$, which reflects the uncertainty predicting a sample to class $c$

$$
d_{\mathcal{M}}^2 = \sum_{c=1}^{C} \left( \frac{1}{N_s} \sum_{i=1}^{N_s} p(y_i^s = c|x_i^s) - \frac{1}{N_t} \sum_{j=1}^{C} p(y_j^t = c|x_j^t) \right)^2.
$$

(37)

Therefore, the conditional-alignment-based deep adaptation model was generally constructed by combining (35) and (37) 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 the source and target samples. Generally, the basic strategies for constraining the joint error $\lambda$ to be small enough refer to as target pseudo-labeling [204], source pretrained network freezing [201], and information theory metric [223], etc.

Very recently, inspired by the privacy protection mechanism, source-free DA models have emerged [235]–[238], in which the source samples are used to pretrain a CNN model but not explicitly exploited in DA.

### C. BN-Layer-Derived

In this section, we introduce another branch, i.e., network layer design by modifying the internal component of CNN for transferable representation. BN has been proven to improve the predictive performance and training efficiency by normalizing the inputs to be zero-mean and univariance, and further scales and shifts the normalized signals with two trainable parameters. Given a batch $B$, each $x_i \in B$ is transformed as $BN(x_i, k) = \gamma_k(x_i - \mu_{B,k})/(\sigma_{B,k} + \epsilon)^{1/2} + \beta_k$, where $k (1 \leq k \leq d)$ indicates the $k$th dimension of $x$, $\mu_{B,k}$ and $\sigma_{B,k}$ are the mean and std. w.r.t. the $k$th dimension in batch $B$, respectively, $\epsilon$ is a constant for numerical stability, and $\gamma_k$ and $\beta_k$ are the learnable scaling and shifting parameters, respectively. For cross-domain applications, due to the datasets’ bias and domain shift, the means and variance in a batch for different domains should be different. Therefore, researchers have proposed to embed domain alignment layers into BN layer for DA [209]–[213].

Roy et al. [211] proposed a domain-specific batch whitening (BW) layer instead of BN for reducing domain shift, which is defined as $BW(x_i, k) = \gamma_k \hat{x}_{i,k} + \beta_k$ and $\hat{x}_i = W_B(x_i - \mu_B)$, such that $W_B W_B^T = \Sigma_B^{-1}$, where $\mu_B$ is the domain-specific mean vector in a batch $B$ and $\Sigma_B^{-1}$ is the domain-specific covariance matrix computed using $B$. Wang et al. [213] proposed a transferable normalization (TransNorm), in which domain-specific mean vectors $\mu_S$ and $\mu_T$ and variances $\sigma_S$ and $\sigma_T$ are computed, while the learnable parameters $\gamma$ and $\beta$ to scale and shift the normalized signal are domain-shared. Similarly, the Auto Dial proposed by Carlucci et al. [210] also shares the two parameters across domains, and Chang et al. [212] also proposed a domain-specific BN layer. Besides BN layer, it is worth to note that Lee [220] proposed a dropout adaptation (DA) via a learnable dropout mask.

### D. Autoencoder-Based

As mentioned above, the training of DNNs needs a large amount of labeled source data. For unsupervised feature learning in DA, a deep-autoencoder-based network adaptation framework was presented [139], [214]–[217], [239]. Generic autoencoders comprise of an encoder function $f(\cdot)$ and a decoder function $g(\cdot)$, which are typically trained to minimize the reconstruction error. Denoising autoencoders (DAEs) were generally constructed with one-layer neural networks for reconstructing the original data from partially or randomly corrupted data [240]. The DAEs can be stacked into a deep network (i.e., SDA), optimized by greedy layer-wise fashion based on stochastic gradient descent (SGD). The rationale behind deep autoencoder-based network adaptation is that the source trained encoder and decoder can be adapted to represent those samples from a target domain.

In [214], Glorot et al. proposed an SDA-based feature representation in conjunction with SVMs for sentiment analysis. Chen et al. [215] proposed a marginalized stacked denoising autoencoder (mSDA), which addressed two crucial limitations of SDAs, high computational cost and low scalability to high-dimensional features, by inducing a closed-form solution without SGD optimization. In [216], 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

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

(38)
where $z$ is the intermediate feature representation of sample $x$. Generally, stacked deep-autoencoder-based TAL framework can be written as

$$\min_{f,g,\theta} J(x, \hat{x}) + \lambda \Omega(z_s, z_t) + \beta \mathcal{L}(y_s, y_t) + \gamma \mathcal{R}(f, g)$$  \tag{39}$$

where $f$ is the domain-shared encoder, $g$ is the domain-shared decoder, $J(\cdot) = J_s(\cdot) + J_t(\cdot)$ represents the reconstruction error loss (e.g., $L_2$-norm squared loss), $\Omega(\cdot)$ is the distribution discrepancy metric between the source feature $z_s$ and the target feature $z_t$, $\mathcal{L}(\cdot)$ is the classification loss (e.g., cross-entropy) with parameter $\theta$ learned on the set $(z_s, y_s)$, and $\mathcal{R}(\cdot)$ is the regularizer of the network parameters of $f$ and $g$. In $[216]$, KL divergence $[241]$-based distribution distance metric was considered. KL is a nonsymmetric measure of the divergence between two probability distributions $P$ and $Q$, which was defined as $D_{KL}(P||Q) = \sum_i P(i) \ln((P(i)/Q(i)))$. A smaller value of $D_{KL}(\cdot)$ means higher similarity of two distributions. Because $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, a symmetric KL version was used in $[216]$, in which $\Omega(\cdot)$ in (39) was written as

$$\Omega(z_s, z_t) = D_{KL}(P_s||P_t) + D_{KL}(P_t||P_s)$$  \tag{40}$$

where $P_t = (\bar{z}_t/\Sigma \bar{z}_t)$ and $P_s = (\bar{z}_s/\Sigma \bar{z}_s)$ represent the distribution of the source and target domains. $\bar{z}_s$ and $\bar{z}_t$ represent the mean vector of encoded feature representations.

Similar to the reconstruction protocol in stacked autoencoder, a related work with deep reconstruction based on convolutional neural networks can be referred to as $[218]$, in which the encoded source feature representation is fed 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 tasks can be learned.

### E. Discussion and Summary

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. 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; and 3) the interpretability of deep adaptation is not optimistic and negative transfer is easily encountered.

## VI. ADVERSARIAL ADAPTATION

Adversarial learning, originated from GAN [37], 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 increasingly popular idea for addressing TAL issues, by minimizing the between-domain discrepancy through an adversarial objective (e.g., domain discriminator), instead of the generic MMD-based domain disparity. In fact, minimizing domain disparity is amount to domain confusion in a learned feature space, where the domain discriminator cannot discriminate which domain a sample comes from. The adversarial-adaptation-based TAL approaches are divided into three types: 1) gradient-reversal-based, 2) minimax-optimization-based, and 3) GAN-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. Actually, DA is in essence a minimax problem. The taxonomy of adversarial adaptation approaches is summarized in Table VI.

### A. 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 [242] first demonstrated that the DA behavior can be achieved by adding a simple but effective gradient reversal layer (GRL). The augmented deep architecture can still be trained using standard SGD-based backpropagation. The gradient-reversal-based adversarial adaptation network consists of three parts: domain-invariant feature representation $\theta_f$, visual classifier $\theta_c$, and domain classifier $\theta_d$. Objectively, $\theta_f$ can be learned by trying to minimize the visual classifier loss $L_c$ and simultaneously maximize the domain classifier loss $L_d$, such that feature representation can be domain-invariant (i.e., domain confusion) and class discriminative. Therefore, in backpropagation optimization of $\theta_f$, the contributed gradients from losses $L_c$ and $L_d$ are $(\partial L_c/\partial \theta_f)$ and $-\lambda (\partial L_d/\partial \theta_f)$, 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 DA $[190]$, $[243]$–$[245]$, $[268]$ under CNN architecture, domain adaptive object detection $[246]$, $[269]$, $[270]$ under Faster-RCNN framework, large-scale kinship verification $[247]$, $[248]$, and fine-grained visual classification $[249]$ under Siamese network. By following a similar protocol with $[242]$, in $[243]$, $[246]$, a binary domain classifier (e.g., source is labeled as +1 and $-1$ for target) was designed as an adversarial objective for learning domain-invariant features by deploying a GRL layer. In $[247]$, $[248]$, two methods, AdvNet and Adv-Kin, were proposed, in which a general Siamese network was constructed with three fully connected $\text{(fc)}$ layers for similarity learning. The reversal gradient with negative multiplier $-\lambda I$ was placed in the 1st $\text{fc}$ layer (MMD loss), the generic contrastive loss was deployed in the 2nd $\text{fc}$ layer, and the softmax guided cross-entropy loss was deployed in the last $\text{fc}$ layer. In $[244]$, Pei et al. argued that single-domain discriminator-based adversarial adaptation only aligns

| Adversarial Adaptation | Model Basis | Reference |
|------------------------|-------------|-----------|
| Gradient Reversal      | Domain confusion & GRL | $[242]$, $[243]$, $[190]$, $[244]$, $[245]$, $[246]$, $[247]$, $[248]$, $[249]$ |
| Minimax Optimization   | Domain confusion & Game | $[250]$, $[251]$, $[252]$ |
| GANs-Based             | Pixel-level Image Synthesis | $[259]$, $[260]$, $[261]$, $[262]$, $[263]$, $[45]$, $[264]$, $[265]$, $[266]$, $[267]$, $[52]$, $[51]$ |
the between-domain distribution without exploiting multimode structures. Therefore, they proposed a multiadversarial 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. [245] proposed a collaborative adversarial network (CAN) by designing multiple domain classifiers, one for each feature extraction block in CNN.

B. Minimax-Optimization-Based

In GANs, the two key parts G and D are often placed with an adversarial state, and generally solved using a minimax-based gaming optimization method [37]. Therefore, the minimax-optimization-based adversarial adaptation can be implemented for domain confusion, through an adversarial objective of the domain discriminator or regressor [208], [250]–[257]. Minimax-optimization-based adversarial adaptation training of DNNs originated in 2015 [250], [251]. Domain confusion maximization-based adversarial DA was first proposed by Tzeng et al. [250], in which an adversarial CNN framework was deployed with classification loss, soft label loss, and two adversarial objectives i.e., domain confusion loss and domain discriminator loss. In [251], Ajakan et al. first 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 $\theta_C$, $\theta_D$, and $\theta_F$. The general adversarial adaptation model aims to minimize the visual classifier loss $L_C$ and maximize the domain discriminator loss $L_D$ by learning $\theta_F$, 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 $L_D$ under $\theta_F$. Generally, maximizing $L_D$ 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:

$$\min_{\theta_C, \theta_D} L_C(D_S, Y_S; \theta_C, \theta_F) - \lambda L_D(D_S, D_T, \theta_D; \theta_F)$$

$$\min_{\theta_D} L_D(D_S, D_T, \theta_F; \theta_D)$$

(41)

where $D_S$ and $D_T$ mean the source and target domain samples, respectively, and $Y_S$ denotes the source data labels.

Under this basic framework in (41), Tzeng et al. [253] further proposed an adversarial discriminative domain adaptation (ADDA) method, in which two CNNs were separately learned for the source and target domains. The training of the source CNN relied only on the source data and labels by minimizing the cross-entropy loss $L_C$, while the target CNN and the domain discriminator loss $L_D$ were alternatively trained in an adversarial fashion with the source CNN fixed. Rozantsev et al. [255] 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 domain-specific feature representation, Long et al. [256] 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. [258] proposed a novel adversarial strategy, i.e., maximum classifier discrepancy (MCD), which aims to maximize the discrepancy between two classifiers’ outputs instead of the 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. In MCD, only two classifiers are designed. By following the wisdom of crowd principle, a stochastic classifier (STAR) [271] which can sample an arbitrary number of classifiers from a Gaussian distribution rather than two classifiers is proposed and achieves better performance. By following the adversarial principle of double classifiers, a classifier competition model for reliable DA was proposed in [272].

C. GAN-Based

In GAN [37] and its variants, two key parts—generator $G$ and discriminator $D$—are generally composed. The generator $G$ aims to synthesize implausible images using the encoder and the 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 the target domain and then enable the cross-domain image classification using synthesized image samples (e.g., objects, scenes, pedestrians, and faces) [45], [259]–[266], [273], [274].

Under the CycleGAN framework proposed by Zhu et al. [44], Hoffman et al. [259] first 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. [260] and Taigman et al. [261] proposed GAN-based models for unsupervised image-level DA, which aims to adapt the source domain images to appear as if drawn from the target domain with well-preserved identity. In [262], 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 domain transformation. Murez et al. [263] and Hong et al. [264] proposed image-to-image translation-based DA models by leveraging GAN and synthetic data for semantic segmentation of the target domain images. Person reidentification (ReID) is typically a cross-domain feature match and retrieval problem [136], [157]. Recently, for addressing ReID challenges in complex scenarios, GAN-based DA was presented for implausible person image generation [265]–[267], across different visual cues and styles, such as poses, backgrounds, lightings, resolutions, and seasons. In addition, GAN-based cross-domain facial image generation for pose-invariant face representation, face frontalization, and rotation were intensively studied [40], [51], [52], [54], all of which tend to address DA across poses for face recognition.
D. Discussion and Summary

In this section, adversarial adaptation is presented with three streams, including gradient reversal, minimax optimization, and GAN. 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; and 3) the gaming between feature generator and discriminator is human-dependent.

VII. BEYOND THE ABOVE: CURRENT FOCUSES IN TAL

We have discussed the new taxonomy from the perspective of general techniques. Consider that TAL is a hot topic and new models are emerging endlessly. To specify the most recent focuses, eight concerns of TAL in this community are summarized, to have a quick clear on very recent progress.

A. Self-Training via Pseudo-Labeling

Due to the unavailability of target labels, pseudo-labeling by setting prediction confidence threshold is often used for self-training. However, pseudo-labeling error accumulation caused by domain shift hampers adaptation. Recent progress focuses on improving the pseudo-label quality by progressive learning from easy to hard [204], class prototype constraints [275], and pseudo-label denoising and densification [276].

B. Multisource DA

Multisource DA is a more realistic than single-source DA. The key challenge is how to exploit the positive but reduce the negative effect of less confident sources. Recent progress focuses on progressively fine-tuning and training source models via selected source samples and features [277], [278], and self-training with target pseudo-labeling by multisource classifiers’ agreement [279].

C. Source-Free and Continual DA

Toward data privacy critical applications, source-free DA is proposed without access to source data [280]. Recent progress focuses on hypothesis transfer, mutual information maximization, and decentralization [62], [237], [238], [281]. Continual DA aims to overcome the forgetting during online learning on evolving domains [112]. Recent progress on continual DA focuses on domain randomization, meta learning, and feature replay [137], [282], [283]. Essentially, for alleviating catastrophic forgetting, learning causal invariance features is crucial toward lifelong learning and DA.

D. Replacement of Entropy Minimization

Due to target label dilemma, with only entropy regularization (minimization), target prediction bias tends to have low class diversity and minority classes of target may be mis-classified as majority classes. Recent progress focuses on improving prediction class diversity and reducing class confusion, by replacing entropy minimization with nuclear norm maximization [284], class confusion minimization [285], etc.

E. Universal DA

Universal DA solves a mixture of closed set, partial set, and open-set DA. Recent progress focuses on weighting common and private classes based on domain similarity and entropy uncertainty assumptions [196], [286], and entropy separation [287].

F. Balancing Alignment and Discrimination

DA is generally a multitask co-training between domain alignment and class discrimination, and inevitably meets imbalance in optimization. Recent progress focuses on between-task gradient consistency maximization by meta learning strategy [288] and dynamic weighting learning between two tasks [289].

G. Domain Generalization

DG is an independent branch of DA, which is more realistic for generalization to unseen target domains. Recent progress focuses on improving source diversity with domain augmentation/randomization [290], learning domain invariance via meta-learning strategy [291], and test time training and inference [292].

H. Adversarial Robustness of DA

DNN is known to be vulnerable to adversarial attacks, and adversarial robustness of DA is largely overlooked. Recent progress on improving adversarial robustness of DA focuses on distillation-based adversarial contrastive losses [293], [294] and adversarial robust training [295].

VIII. APPLICATION OF TAL IN OTHER VISION TASKS

In this section, from the perspective of applications, we have a discussion on TAL theory used in other fundamental vision tasks, such as object detection and semantic segmentation, as illustrated in Fig. 1. It is not difficult to imagine that, for example, fog caused-image degradation will inevitably lead to detection performance degradation. A preliminary research in [296] finds that degradation removal does not help CNN-based image classification, because degradation removal focuses on pixel manipulation for image enhancement but does not bring any new information beneficial to classification. Therefore, cross-dataset object detection/segmentation (e.g., from cityscapes to foggy cityscapes) rather than learning to dehaze is an important but challenging vision task. In the past two years, this topic has attracted researchers’ attention and a significant progress is witnessed. Specifically, many domain adaptive object detection models [246], [269], [270], [297]–[301] and domain adaptive semantic segmentation models [302]–[309] have been proposed. These works have well-revealed that the versatile TAL theory and algorithm can be combined with detection network (e.g., Faster-RCNN) or segmentation network (i.e., CNN-based) for more generalized vision tasks in the wild. Besides, VTAL has also been used for cross-domain/cross-modal visual retrieval [310]–[312], person reidentification [313], person search [314], and image restoration and dehazing [315], [316]. It is hopeful that VTAL will play an increasingly important role in more vision tasks with more heuristic models and algorithms toward transferable and universal representations.

A. SOTA Methods on Classification Benchmarks

The standard protocol of TAL approaches is measuring the image classification accuracy on unlabeled target domain. As far as we know, 14 benchmarks for image classification are used. Due to space limitation, 61 methods on 10 classification benchmarks are summarized in Supplementary Materials.
TAL 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 article surveys recent advances in TAL in the past decade and presents a new taxonomy of five technical challenges being faced by researchers: instance reweighting adaptation, feature adaptation, classifier adaptation, deep network adaptation, and adversarial adaptation. Actually, a number of models are crossed among them, and the most appealing combination in most recent works is reweighting + deep network + adversarial strategies, which are not specially discussed due to space limitation.

The proposed taxonomy of TAL provides a framework for better understanding and identifying the status, future research challenges, and directions. Each challenge is 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. The typical problems of TL and DA are negative transfer (i.e., model over-fitting) and underadaptation (i.e., model underfitting), which attracted many researchers’ focus in the past decade, but not well-addressed due to the difficulty in explicitly characterizing the domain discrepancy. There are lacks of some judgment conditions and evaluation criteria. Negative transfer can be stated as transferring knowledge from source can have a negative impact on the target learner. In [317], three impacting factors leading to underlying negative transfer are stressed, such as algorithmic choice, distribution divergence, and the size of labeled target data. In addition, the essence of TAL is to learn domain-invariant but class-discriminative representation for unlabeled target data regardless of close set, open set, or partial DA. Therefore, unsupervised/self-supervised learning idea can be a choice toward more generalizable representations beyond domain-invariant representations.

Throughout the entire research lines, one specific research area of TAL that seems to be still understudied is domain generalization to unseen (out-of-distribution) domains. This challenge is more approaching real-world scenarios that numerous domains can be found [318], which is also related to continual/lifelong learning with continuous domains. Domain generalization expects to capture commonality (invariance). Learning causality invariance can be a potential solution by combining causal learning with TL. To uncover the knowledge used to transfer, combining domain augmentation and meta learning that learns to learn can be a choice. The meta-TL in [319] and task disentanglement [320] show some instructive insights on learning generalizable representations. Another branch with little attention is the adversarial robustness of TAL models due to the vulnerability to human-imperceptible attacks. Therefore, robust DA theory under known/unknown adversarial attacks needs to be studied. Note that this survey focuses on DA, and zero/few-shot learning is not included in this work.

Undoubtedly, a huge and pleasant progress in TAL model and algorithm is witnessed with blowout publications, which has well-answered how to facilitate DA under the well-established expected target error upper bound theory. However, we are not optimistic about this, and three open questions are stressed. 1) When will we need TAL for a given application scenario? The basic analyzing condition of whether cross-domain happens is still not clear, which makes TAL blind. 2) Where does the DA theory go when two domains are completely disjoint (i.e., large domain discrepancy) in open-world (dynamic) scenarios? The DA theory is scenario-dependent and the impossibility theory [321] of TAL needs to be broken. Also, studying the target error lower bound may be more instructive than the expected error upper bound, but seems to be paid less attention. 3) What useful knowledge has been transferred from the source to the target? This has been preliminarily explained and studied in [322], [323]. We observe these heuristic but promising directions of TAL for future research. Addressing these issues and challenges can well improve the universality, interpretability, and credibility of TAL in open-world scenarios toward safer applications.

REFERENCES

[1] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” in Proc. CVPR, Jun. 2011, pp. 1521–1528.
[2] H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” J. Statist. Planning 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 Proc. ACM MM, 2007, pp. 188–197.
[4] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, Oct. 2009.
[5] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, “Adapting visual category models to new domains,” in Proc. ECCV, 2010, pp. 213–226.
[6] L. Pratt, J. Mostow, and C. Kanam, “Direct transfer of learned information among neural networks,” in Proc. AAAI, 1991, pp. 584–589.
[7] L. Pratt, “Discriminability-based transfer between neural networks,” in Proc. NIPS, 1993, pp. 1–8.
[8] S. M. Salaken, A. Khosravi, T. Nguyen, and S. Nahavandi, “ extreme learning machine based transfer learning algorithms: A survey,” Neurocomputing, vol. 267, pp. 516–524, Dec. 2017.
[9] K. Weiss, T. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” J. Big Data, vol. 3, no. 1, pp. 1–40, 2016.
[10] L. Shao, F. Zhu, and X. Li, “Transfer learning for visual categorization: A survey,” IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 5, pp. 1019–1034, May 2014.
[11] W. Pan, “A survey of transfer learning for collaborative recommendation with auxiliary data,” Neurocomputing, vol. 177, pp. 447–453, Feb. 2016.
[12] V. Patel, R. Gopalan, R. Li, and R. Chellappa, “Visual domain adaptation: A survey of recent advances,” IEEE Signal Process. Mag., vol. 32, no. 3, pp. 53–69, May 2015.
[13] J. Li, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, “Transfer learning using computational intelligence: A survey,” Knowl-Based Syst., vol. 80, pp. 14–23, May 2015.
[14] M. Wang and W. Deng, “Deep visual domain adaptation: A survey,” Neurocomputing, vol. 312, pp. 135–153, Oct. 2018.
[15] D. Cook, K. D. Feuz, and N. C. Krishnan, “Transfer learning for activity recognition: A survey,” Knowl. Inf. Syst., vol. 36, no. 3, pp. 537–556, Sep. 2013.
[16] G. Csurka, “Domain adaptation for visual applications: A comprehensive survey,” 2017, arXiv:1702.05374.
[17] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A survey on deep transfer learning,” 2018, arXiv:1808.01974.
[18] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504–507, 2006.
[19] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, Sep. 2015.
[20] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. NIPS, 2012, pp. 1–9.
[21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. CVPR, Jun. 2016, pp. 770–778.
[22] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. NIPS, 2015, pp. 1–9.
et al.  [37] I. J. Goodfellow, Y. Bengio, “Deep learning of representations for unsupervised and supervised learning,” in Proc. NIPS, 2016, pp. 1–8.

[36] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in Proc. ICLR, 2016, pp. 1–16.

[35] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. CVPR, 2017, pp. 1125–1134.

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. ICCV, Oct. 2017, pp. 2223–2232.

[33] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation,” in Proc. CVPR, Jun. 2018, pp. 8789–8797.

[32] M. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” in Proc. NIPS, 2017, pp. 1–9.

[31] D. Tao, N. King, A. Paek, and I. Kweon, “Pixel-level domain transfer,” in Proc. ECCV, 2016, pp. 517–532.

[30] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in Proc. CVPR, Jun. 2016, pp. 2414–2423.

[29] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, “Controlling perceptual factors in neural style transfer,” in Proc. CVPR, Jul. 2017, pp. 9097–9097.

[28] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. ECCV, 2016, pp. 694–711.

[27] D. Chung, K. Tahboub, and E. J. Delp, “A two stream Siamese convolutional neural network for person re-identification,” in Proc. ICCV, Oct. 2017, pp. 1953–1991.

[26] 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 Proc. CVPR, Jun. 2016, pp. 1335–1344.

[25] L. Hou, D. Samaras, T. M. Kure, Y. Gao, J. E. Davis, and J. H. Saltz, “Patch-based convolutional neural network for whole slide tissue image classification,” in Proc. CVPR, Jun. 2016, pp. 2424–2433.

[24] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, no. 7639, pp. 115–118, 2017.

[23] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: Object detection via region-based fully convolutional networks,” in Proc. NIPS, 2016, pp. 1–9.

[22] W. Liu et al., “SSD: Single shot multibox detector,” in Proc. ECCV, 2016, pp. 21–37.

[21] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, “DDSSD: Deconvolutional single shot detector,” 2017, arXiv:1701.06659.

[20] E. Ahmed, M. Jones, and T. K. Marks, “An improved deep learning architecture for person re-identification,” in Proc. CVPR, Jun. 2015, pp. 3908–3916.

[19] D. Chung, K. Tahboub, and E. J. Delp, “A two stream Siamese convolutional neural network for person re-identification,” in Proc. ICCV, Oct. 2017, pp. 1953–1991.

[18] 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 Proc. CVPR, Jun. 2016, pp. 1335–1344.

[17] L. Hou, D. Samaras, T. M. Kure, Y. Gao, J. E. Davis, and J. H. Saltz, “Patch-based convolutional neural network for whole slide tissue image classification,” in Proc. CVPR, Jun. 2016, pp. 2424–2433.

[16] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol. 542, no. 7639, pp. 115–118, 2017.

[15] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: Object detection via region-based fully convolutional networks,” in Proc. NIPS, 2016, pp. 1–9.

[14] W. Liu et al., “SSD: Single shot multibox detector,” in Proc. ECCV, 2016, pp. 21–37.

[13] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, “DDSSD: Deconvolutional single shot detector,” 2017, arXiv:1701.06659.

[12] E. Ahmed, M. Jones, and T. K. Marks, “An improved deep learning architecture for person re-identification,” in Proc. CVPR, Jun. 2015, pp. 3908–3916.
H. Xu, J. Zheng, and R. Chellappa, “Bridging the domain shift by domain adaptive dictionary learning,” in IEEE Trans. BMVC, 2015, p. 1.

S. Shekhar, V. M. Patel, H. Van Nguyen, and R. Chellappa, “Coupled projections for adaptation of dictionaries,” IEEE Trans. Image Process., vol. 24, no. 10, pp. 2941–2954, Oct. 2015.

Q. Qiu and R. Chellappa, “Compositional dictionaries for domain adaptive face recognition,” IEEE Trans. Image Process., vol. 24, no. 12, pp. 5152–5165, Dec. 2015.

Q. Qiu, V. Patel, T. Furuta, and R. Chellappa, “Domain adaptive metric learning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 11, pp. 2642–2656, Nov. 2013.

J. Ni, Q. Qiu, and R. Chellappa, “Subspace interpolation via dictionary learning for unsupervised domain adaptation,” in Proc. CVPR, Jun. 2013, pp. 692–699.

F. Zhu and L. Shao, “Weakly-supervised cross-domain dictionary learning for visual recognition,” Int. J. Comput. Vis., vol. 109, nos. 1–2, pp. 42–59, Aug. 2014.

B. Lu, R. Chellappa, and N. M. Nasrabadi, “Incremental dictionary learning for unsupervised domain adaptation,” in Proc. BMVC, 2015, pp. 108.1–108.12.

J. Hoffman, T. Darrell, and K. Saenko, “Continuous manifold based adaptation for evolving visual domains,” in Proc. CVPR, Jun. 2014, pp. 867–874.

B. Kulis, K. Saenko, and T. Darrell, “What you saw is not what you get: Domain adaptation using asymmetric kernel transforms,” in Proc. CVPR, Jun. 2011, pp. 1785–1792.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, “Optimal transport for domain adaptation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 9, pp. 1853–1865, Sep. 2017.

M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann, “Unsupervised domain adaptation by domain invariant projection,” in Proc. ICCV, Dec. 2013, pp. 769–776.

M. Long, J. Wang, I. W. Tsang, and P. S. Yu, “Domain invariant transfer kernel learning,” IEEE Trans. Knowl. Data Eng., vol. 27, no. 6, pp. 1519–1532, Jun. 2014.

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,” Sens. Actuators B, Chem., vol. 253, pp. 407–417, Dec. 2017.

V. Rezaa, D. Balduzzi, W. B. Kleijn, and M. Zhang, “Scatter component analysis: A unified framework for domain adaptation and domain generalization,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 7, pp. 1414–1430, Jul. 2017.

J. Zhang, W. Li, and P. Ogumbona, “Joint geometrical and statistical alignment for visual domain adaptation,” in Proc. CVPR, Jul. 2017, pp. 1859–1867.

L. Zhang, Y. Liu, and P. Deng, “Ood recognition in multiple E-nose systems with cross-domain discriminative subspace learning,” IEEE Trans. Instrum. Meas., vol. 66, no. 7, pp. 1679–1692, Jul. 2017.

H. Lu, C. Shen, Z. Cao, Y. Xiao, and A. van den Hengel, “An embarrassing simple approach to visual domain adaptation,” IEEE Trans. Image Process., vol. 27, no. 7, pp. 3403–3417, Jul. 2018.

S. Li, S. Song, G. Huang, Z. Ding, and C. Wu, “Domain invariant and class discriminative feature learning for visual domain adaptation,” IEEE Trans. Image Process., vol. 28, no. 9, pp. 4269–4273, Sep. 2019.

S. Li, S. Song, G. Huang, and C. Wu, “Cross-domain extreme learning machines for domain adaptation,” IEEE Trans. Syst., Man, Cybern., Syst., vol. 49, no. 6, pp. 1194–1207, Jun. 2019.

A. Gretton, O. Bousquet, A. Smola, and B. Scholkopf, “Measuring statistical dependence with Hilbert–Schmidt norms,” in Proc. ALT, 2005, pp. 63–77.

K. Yan, L. Kou, and D. Zhang, “Learning domain-invariant subspace using domain features and independence maximization,” IEEE Trans. Cybern., vol. 48, no. 1, pp. 288–299, Jan. 2018.

Y. Zhang and D.-Y. Yeung, “Transfer metric learning by learning task relationships,” in Proc. ACM SIGKDD, 2010, pp. 1199–1208.

J. Hu, J. Lu, Y.-P. Tan, and J. Zhou, “Deep transfer metric learning,” IEEE Trans. Image Process., vol. 25, no. 12, pp. 5576–5588, Dec. 2016.

B. Geng, D. Tao, and C. Xu, “DAML: Domain adaptation metric learning,” IEEE Trans. Image Process., vol. 20, no. 10, pp. 2980–2989, Oct. 2011.

Y. Xu et al., “A unified framework for metric transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 6, pp. 1158–1171, Jun. 2017.

S. Li et al., “Discriminative transfer feature and label consistency for cross-domain image classification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 11, pp. 4542–4556, Sep. 2020.

B. Sun, J. Feng, and K. Saenko, “Return of frustratingly easy domain adaptation,” in Proc. AAAI, 2016, pp. 153–171.
M. Long, J. Wang, and M. I. Jordan, “Unsupervised domain adaptation,” in Proc. CVPR, Jun. 2015, pp. 1–9.

Z. Cao, M. Long, J. Wang, and M. I. Jordan, “Partial transfer learning with selective adversarial networks,” in Proc. CVPR, Jun. 2018, pp. 2724–2733.

J. Zhang, Z. Ding, W. Li, and P. Oggunbona, “Importance weighted adversarial nets for partial domain adaptation,” in Proc. CVPR, Jun. 2018, pp. 8156–8164.

S. Li et al., “Deep residual correction network for partial domain adaptation,” in Proc. CVPR, Jun. 2020, pp. 2985–2994.

Y. Pan, T. Yao, Y. Li, C.-W. Ngo, and T. Mei, “Exploring category-agnostic clusters for open-set domain adaptation,” in Proc. CVPR, Jun. 2020, pp. 13867–13875.

J. N. Kundu, N. Venkat, A. Revanur, M. V. Rahul, and R. V. Babu, “Towards inheritable models for open-set domain adaptation,” in Proc. CVPR, Jun. 2020, pp. 12376–12385.

Z. Fang, J. Lu, F. Liu, J. Xuan, and G. Zhang, “Open set domain adaptation: Theoretical bound and algorithm,” IEEE Trans. Neural Netw. Learn. Syst., vol. 43, no. 7, pp. 4309–4322, Oct. 2021.

Z. Cao, K. You, M. Long, J. Wang, and Q. Yang, “Learning to transfer examples for partial domain adaptation,” in Proc. CVPR, Jun. 2019, pp. 4119–4128.

Y. Pan, T. Yao, Y. Li, and P. He, “Supervised representation learning: Transfer learning with deep autoencoders,” in Proc. IJCAI, 2015, pp. 4119–4125.

L. Wen, L. Gao, and X. Li, “A new deep transfer learning based on sparse auto-encoder for fault diagnosis,” IEEE Trans. Syst., Man, Cybern., Syst., vol. 49, no. 1, pp. 136–144, Jan. 2019.

M. Zhihary, W. B. Kleijn, D. Baldiuzzi, and W. Li, “Deep reconstruction-classification networks for unsupervised domain adaptation,” in Proc. ECCV, 2016, pp. 597–613.

Y. Pan, T. Yao, Y. Li, and P. He, “Towards prototype-based unsupervised domain adaptation,” in Proc. CVPR, Jun. 2019, pp. 2239–2247.

S. Lee, D. Kim, N. Kim, S. G. Jeong, “Drop to adapt: Learning discriminative features for unsupervised domain adaptation,” in Proc. ICCV, Oct. 2019, pp. 91–100.

H. Tang, K. Chen, and K. Jia, “Unsupervised domain adaptation via structurally regularized deep clustering,” in Proc. CVPR, Jun. 2020, pp. 8725–8735.

Y.-C. Liu, Y.-Y. Yeh, T.-C. Fu, S.-D. Wang, W.-C. Chiu, and Y.-C.-F. Wang, “Detach and adapt: Learning cross-domain disentangled deep representation,” in Proc. CVPR, Jun. 2018, pp. 8867–8876.

H. Zhao, R. T. D. Combes, K. Zhang, and G. Gordon, “On learning invariant representations for domain adaptation,” in Proc. ICMIL, 2019, pp. 7523–7532.

S. Yu, J. Tzeng, T. Batra, M. H. Baig, and D. Ullrich, “Sliced Wasserstein discrepancy for unsupervised domain adaptation,” in Proc. CVPR, Jun. 2019, pp. 10285–10295.

Y. Balaji, R. Chellappa, and S. Feizi, “Normalized Wasserstein for mixture distributions with applications in adversarial learning and domain adaptation,” in Proc. IJCV, Oct. 2019, pp. 6500–6508.

R. Xu, G. Li, J. Yang, and L. Lin, “Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation,” in Proc. ICCV, Oct. 2019, pp. 1426–1435.

Y. Zhang, T. Liu, M. Long, and M. Jordan, “Bridge theory and algorithm for domain adaptation,” in Proc. IJCAI, 2019, pp. 7404–7413.

X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, “Model matching for multi-source domain adaptation,” in Proc. ICCV, Oct. 2019, pp. 1406–1415.

M. Li, Y.-M. Zhai, Y.-W. Luo, P.-F. Ge, and C.-X. Ren, “Enhanced transport distance for unsupervised domain adaptation,” in Proc. CVPR, Jun. 2020, pp. 13936–13944.

R. Xu, P. Liu, L. Wang, C. Chen, and J. Wang, “Reliable weighted optimal transport for unsupervised domain adaptation,” in Proc. CVPR, Jun. 2020, pp. 4394–4403.

Y. Zhang, F. Liu, Z. Fang, B. Yuan, G. Zhang, and J. Lu, “Learning from a complementary-label source domain: Theory and algorithms,” IEEE Trans. Neural Netw. Learn. Syst., early access, Jun. 17, 2021, doi: 10.1109/TNNLS.2021.3089099.

Y. Zhang, F. Liu, Z. Fang, B. Yuan, G. Zhang, and J. Lu, “ClariNet: A one-step approach towards budget-friendly unsupervised domain adaptation,” in Proc. IJCAI, Jul. 2020.

Y. Grandvalet and Y. Bengio, “Semi-supervised learning by entropy regularization,” in Proc. NIPS, 2004, pp. 529–536.

H. Liu, M. Long, J. Wang, and M. Jordan, “Transferable adversarial training: A general approach to adapting deep classifiers,” in Proc. ICML, 2019, pp. 4013–4022.

R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu, “Model adaptation: Unsupervised domain adaptation without source data,” in Proc. CVPR, Jun. 2020, pp. 9641–9650.

J. N. Kundu, N. Venkat, M. V. Rahul, and R. V. Babu, “Universal source-free domain adaptation,” in Proc. CVPR, Jun. 2020, pp. 4544–4553.

S. Kullback, “Letter to editor: The Kullback–Leibler distance,” Tech. Rep., 1987.
Z. Lu, Y. Yang, X. Zhu, C. Liu, Y.-Z. Song, and T. Xiang, “Stochastic classifiers for unsupervised domain adaptation,” in Proc. CVPR, Jun. 2020, pp. 9111–9120.

J. Fu and L. Zhang, “Reliable domain adaptation with classifiers competition for image classification,” IEEE Trans. Circuits Syst. II, Exp. Briefs, vol. 67, no. 11, pp. 2802–2806, Nov. 2020.

S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, “Generate to adapt: Aligning domains using generative adversarial networks,” in Proc. CVPR, Jun. 2018, pp. 8503–8512.

R. Xu, Z. Chen, W. Zuo, J. Yan, and L. Lin, “Deep cocktail network: Multi-source unsupervised domain adaptation with category shift,” in Proc. CVPR, Jun. 2018, pp. 3964–3973.

J. Liang, D. Hu, and J. Feng, “Domain adaptation with auxiliary target domain-oriented classifier,” in Proc. CVPR, Jun. 2021, pp. 16632–16642.

I. Shin, S. Woo, F. Pan, and I. Kweon, “Two-phase pseudo label den- sification for self-training based domain adaptation,” in Proc. ECCV, 2020, pp. 532–548.

S. Zhao et al., “Multi-source distilling domain adaptation,” in Proc. AAAI, 2020, pp. 12975–12983.

Y. Fu et al., “Partial feature selection and alignment for multi-source domain adaptation,” in Proc. CVPR, Jun. 2021, pp. 16654–16663.

N. Venkat, J. Kundu, D. Singh, A. Revanur, and V. R. Babu, “Your classifier can secretly suffice multi-source domain adaptation,” in Proc. NeurIPS, 2020, pp. 1–13.

B. Chidlovskii, S. Clinchant, and G. Csurka, “Domain adaptation in the absence of source data domain,” in Proc. ACM SIGKDD, Aug. 2016, pp. 451–460.

S. M. Ahmed, D. S. Raychaudhuri, S. Paul, S. Oymak, and A. K. Roy-Chowdhury, “Unsupervised multi-source domain adaptation without access to source data,” in Proc. CVPR, Jun. 2021, pp. 10103–10112.

R. Volpi, D. Larlus, and G. Rogez, “Continual adaptation of visual representations via domain randomization and meta-learning,” in Proc. CVPR, Jun. 2021, pp. 4443–4453.

H. Wang, H. He, and D. Katabi, “Continuously indexed domain adaptation,” in Proc. ICML, 2020, pp. 9898–9907.

S. Cui, S. Wang, J. Zhao, L. Li, Q. Huang, and Q. Tian, “Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations,” in Proc. CVPR, Jun. 2020, pp. 3941–3950.

Y. Jin, X. Wang, M. Long, and J. Wang, “Minimum class confusion for versatile domain adaptation,” in Proc. ECCV, 2020, pp. 464–480.

Q. Yu, A. Hashimoto, and Y. Ushiku, “Divergence optimization for noisy universal domain adaptation,” in Proc. CVPR, Jun. 2021, pp. 2515–2524.

K. Saito, D. Kim, and S. Sclaroff, “Universal domain adaptation through self-supervision,” in Proc. NeurIPS, 2020, pp. 1–11.

G. Wei, C. Lan, W. Zeng, and Z. Chen, “MetaAlign: Coordinating domain alignment and classification for unsupervised domain adaptation,” in Proc. CVPR, Jun. 2021, pp. 16643–16653.

N. Xiao and L. Zhang, “Dynamic weighted learning for unsupervised domain adaptation,” in Proc. CVPR, Jun. 2021, pp. 15242–15251.

J. Huang, D. Guan, A. Xiao, and S. Lu, “FSDR, Frequency space domain randomization for domain generalization,” in Proc. CVPR, Jun. 2021, pp. 6891–6902.

Y. Shu, Z. Cao, C. Wang, J. Wang, and M. Long, “Open domain generalization with domain-augmented meta-learning,” in Proc. CVPR, Jun. 2021, pp. 9624–9633.

P. Pandey, M. Raman, S. Varambally, and A. Prathosh, “Generalization on unseen domains via inference-time label-preserving target projections,” in Proc. CVPR, 2021, pp. 12924–12933.

J. Zhang, H. Chao, and P. Yan, “Robustified domain adaptation,” 2020, arXiv:2011.09563.

J. Yang et al., “Exploring robustness of unsupervised domain adaptation in semantic segmentation,” in Proc. ICCV, Oct. 2021, pp. 9194–9203.

L. Zhang, Y. Zhou, and L. Zhang, “On the robustness of domain adaptation to adversarial attacks,” 2021, arXiv:2108.01807.

Y. Pei, Y. Huang, Q. Zou, X. Zhang, and S. Wang, “Effects of image degradation and degradation removal to CNN-based image classification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 4, pp. 1239–1253, Apr. 2021.

K. Saito, Y. Ushiku, T. Harada, and K. Saenko, “Strong-weak distribution alignment for adaptive object detection,” in Proc. CVPR, Jun. 2019, pp. 6956–6965.

X. Zhu, J. Pang, C. Yang, J. Shi, and D. Lin, “Adapting object detectors via selective cross-domain alignment,” in Proc. CVPR, Jun. 2019, pp. 687–696.
C.-D. Xu, X.-R. Zhao, X. Jin, and X.-S. Wei, “Exploring categorical regularization for domain adaptive object detection,” in Proc. CVPR, Jun. 2020, pp. 11724–11733.

C. Chen, Z. Zheng, X. Ding, Y. Huang, and Q. Dou, “Harmonizing transferability and discriminability for adapting object detectors,” in Proc. CVPR, Jun. 2020, pp. 8869–8878.

Y. Zheng, D. Huang, S. Liu, and Y. Wang, “Cross-domain object detection through coarse-to-fine feature adaptation,” in Proc. CVPR, Jun. 2020, pp. 13766–13775.

T. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, “Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation,” in Proc. CVPR, 2019, pp. 2517–2526.

Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, “Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation,” in Proc. CVPR, Jun. 2019, pp. 2507–2516.

Y. Li, L. Yuan, and N. Vasconcelos, “Bidirectional learning for domain adaptation of semantic segmentation,” in Proc. CVPR, Jun. 2019, pp. 6936–6945.

Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, “CfDoCo: Pixel-level domain transfer with cross-domain consistency,” in Proc. CVPR, Jun. 2019, pp. 1791–1800.

T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Perez, “DADA: Depth-aware domain adaptation in semantic segmentation,” in Proc. ICCV, Oct. 2019, pp. 7364–7373.

M. Chen, H. Xue, and D. Cai, “Domain adaptation for semantic segmentation with maximum squares loss,” in Proc. ICCV, Oct. 2019, pp. 2000–2009.

F. Lv, T. Liang, X. Chen, and G. Lin, “Cross-domain semantic segmentation via domain-invariant interactive relation transfer,” in Proc. CVPR, Jun. 2020, pp. 4334–4343.

Y. Zhang, Z. Qiu, T. Yao, C.-W. Ngo, D. Liu, and T. Mei, “Transferring and regularizing prediction for semantic segmentation,” in Proc. CVPR, Jun. 2020, pp. 9621–9630.

F. Huang, L. Zhang, Y. Yang, and X. Zhou, “Probability weighted compact feature for domain adaptive retrieval,” in Proc. CVPR, Jun. 2020, pp. 9582–9591.

X. Huang, Y. Peng, and M. Yuan, “MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval,” IEEE Trans. Cybern., vol. 50, no. 3, pp. 1047–1059, Mar. 2020.

D. Kim et al., “Learning cross-modal contrastive features for video domain adaptation,” in Proc. ICCV, Oct. 2021, pp. 13618–13627.

Y. Zhai et al., “AD-cluster: Augmented discriminative clustering for domain adaptive person re-identification,” in Proc. CVPR, Jun. 2020, pp. 9021–9030.

Y. Jing, W. Wang, L. Wang, and T. Tan, “Cross-modal cross-domain moment alignment network for person search,” in Proc. CVPR, Jun. 2020, pp. 10676–10686.

B. Yan, C. Ma, B. Bare, W. Tan, and S. Hoi, “Disparity-aware domain adaptation in stereo image restoration,” in Proc. CVPR, Jun. 2020, pp. 13179–13187.

Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang, “Domain adaptation for image dehazing,” in Proc. CVPR, Jun. 2020, pp. 2808–2817.

Z. Wang, Z. Dai, B. Poczos, and J. Carbonell, “Characterizing and avoiding negative transfer,” in Proc. CVPR, Jun. 2019, pp. 11293–11302.

M. Mancini, L. Porzi, S. R. Bulo, B. Caputo, and E. Ricci, “Boosting domain adaptation by discovering latent domains,” in Proc. CVPR, Jun. 2018, pp. 3771–3780.

Y. Jang, H. Lee, S. Hwang, and J. Shin, “Learning what and where to transfer,” in Proc. ICML, 2019, pp. 3030–3039.

A. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese, “Taskonomy: Disentangling task transfer learning,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3712–3722.

S. Ben-David, T. Luu, T. Lu, and D. Pal, “Impossibility theorems for domain adaptation,” in Proc. AISTATS, 2010, pp. 129–136.

J. Dong, Y. Cong, G. Sun, Z. Fang, and Z. Ding, “Where and how to transfer: Knowledge aggregation-induced transferability perception for unsupervised domain adaptation,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Nov. 16, 2021, doi: 10.1109/TPAMI.2021.3128560.

J. Dong, Y. Cong, G. Sun, B. Zhong, and X. Xu, “What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation,” in Proc. CVPR, Jun. 2020, pp. 4023–4032.