Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective

Jing Zhang  
*University of Wollongong, jz960@uowmail.edu.au*

Wanqing Li  
*University of Wollongong, wanqing@uow.edu.au*

Philip O. Ogunbona  
*University of Wollongong, philipo@uow.edu.au*

Dong Xu  
*University of Sydney*

Publication Details  
Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective

Abstract
This article takes a problem-oriented perspective and presents a comprehensive review of transfer-learning methods, both shallow and deep, for cross-dataset visual recognition. Specifically, it categorises the cross-dataset recognition into 17 problems based on a set of carefully chosen data and label attributes. Such a problem-oriented taxonomy has allowed us to examine how different transfer-learning approaches tackle each problem and how well each problem has been researched to date. The comprehensive problem-oriented review of the advances in transfer learning with respect to the problem has not only revealed the challenges in transfer learning for visual recognition but also the problems (e.g., 8 of the 17 problems) that have been scarcely studied. This survey not only presents an up-to-date technical review for researchers but also a systematic approach and a reference for a machine-learning practitioner to categorise a real problem and to look up for a possible solution accordingly.

Disciplines
Engineering | Science and Technology Studies

Publication Details

This journal article is available at Research Online: https://ro.uow.edu.au/eispapers1/2465
Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective

JING ZHANG, WANQING LI, and PHILIP OGUNBONA, University of Wollongong, Australia
DONG XU, University of Sydney, Australia

This paper takes a problem-oriented perspective and presents a comprehensive review of transfer learning methods, both shallow and deep, for cross-dataset visual recognition. Specifically, it categorizes the cross-dataset recognition into seventeen problems based on a set of carefully chosen data and label attributes. Such a problem-oriented taxonomy has allowed us to examine how different transfer learning approaches tackle each problem and how well each problem has been researched to date. The comprehensive problem-oriented review of the advances in transfer learning with respect to the problem has not only revealed the challenges in transfer learning for visual recognition, but also the problems (e.g. eight of the seventeen problems) that have been scarcely studied. This survey not only presents an up-to-date technical review for researchers, but also a systematic approach and a reference for a machine learning practitioner to categorize a real problem and to look up for a possible solution accordingly.

1 INTRODUCTION

Humans have exceptional ability to transfer learning in one context to another context [178, 246]. Machine learning algorithms mostly inspired by human brains, however, usually require a huge amount of training examples to learn a new model from scratch and often fail to apply the learned model to test data acquired from the scenarios different from those of the training data mainly due to domain divergence and task divergence [170]. This is particularly true in visual recognition [223] where external factors such as environments, lighting, background, sensor types, view angles, and post-processing can cause the distribution shift or even feature space divergence of the same task in two datasets or even the tasks, i.e. categories of the objects, are different.

To use previously available data effectively for current tasks with scarce data, models or knowledge learned from one domain have to be transferred to a new domain for the current task. Transfer learning has been actively researched in the past decade and one of its topics, domain adaptation, has been especially extensively researched, where the previous and current tasks are the same. The extensive study has led to about a dozen of tutorial and survey papers published since 2009, from the analysis of the nature of dataset shift [184] to the formal definition and task-oriented categorization of transfer learning [170], and to the recent tutorial and survey on deep learning based domain adaptation [37, 228]. Most of these survey papers [12, 35, 154, 157, 160, 173, 195, 214, 228, 245] are method-driven and provide up to the time a review of the evolution of the technologies. Many of them are on particular topics, for instance, domain adaptation [12, 37, 157, 173, 214, 228], dataset shift [160], activity recognition [35], and speech and language processing [237]. While these review papers have provided researchers in the field valuable references and contributed significantly to the advances of the technologies, they have not examined the full landscape of transfer learning and maturity of technologies to serve as a reference for machine learning practitioners. Unlike these existing survey papers, this paper takes a new problem-oriented perspective and presents a comprehensive review of transfer learning methods for cross-dataset visual recognition. Specifically,
• It defines a set of data and label attributes, categorises in a fine-grained way the cross-dataset recognition into seventeen problems based on these attributes, and presents a comprehensive review of the transfer learning methods, both shallow and deep, developed to date for each problem.
• The paper has also provided an assessment of the suitability of widely used datasets for transfer learning in evaluating algorithms for each of the seventeen problems.
• The problem-oriented taxonomy has allowed us to examine how different transfer learning approaches tackle each problem, how well each problem has been studied to date and the available solutions to each problem.
• Through the problem-oriented analysis, challenges and future directions have been identified. Particularly, little studies have been reported on eight of the seventeen problems.
• This survey not only presents an up-to-date technical review for researchers, but also a systematic approach and a reference for a machine learning practitioner to categorise a real problem and to look up for a possible solution accordingly.

In addition, none of the previous survey papers covers all of the seventeen problems. For instance, Weiss et al. [245] focuses on nine (of the seventeen) problems on homogeneous and heterogeneous domain adaptation and transfer learning with heterogeneous label spaces; Venkateswara et al. [229] mainly reviewed the literature of two problems in homogeneous domain adaptation using deep-learning; and Csurka [37] focuses on seven problems in domain adaptation.

The rest of the paper is organised as follows. Section 2 explains the terminologies used in the paper, defines the problem-oriented taxonomy of cross-dataset recognition, and summarises the transfer learning approaches to cross-dataset recognition. The seventeen problems identified in the taxonomy are categorised into four scenarios: homogeneous feature and label spaces, heterogeneous feature spaces, heterogeneous label spaces and heterogeneous feature and label spaces. Sections 3 through 6 review and analyse respectively the advances of techniques in addressing the problems under the four scenarios. Section 7 discusses and examines the suitability of the most commonly used datasets for cross-dataset transfer learning for all the problems. Section 8 discusses the challenges and future research directions. Section 9 concludes the paper.

2 OVERVIEW

This section begins with the definitions of terminologies used throughout the paper and then provides a summary of the approaches that have been developed for transfer learning.

2.1 Terminologies and Definitions

In this paper, we follow the definitions of “domain” and “task” given by [170].

Definition 2.1. (Domain [170]) “A domain is defined as \( D = \{X, P(x)\} \), which is composed of two components: a feature space \( X \) and a marginal probability distribution \( P(x) \), where \( x \in X \).”

Definition 2.2. (Task [170]) “Given a specific domain, a task is defined as \( T = \{Y, f(x)\} \), which is composed of two components: a label space \( Y \) and a predictive function \( f(x) \), where \( f(x) \) can be seen as a conditional distribution \( P(y|x) \) and \( y \in Y \).”

Definition 2.3. (Dataset) A dataset is defined as \( S = \{N, X, P(x), Y, f(x)\} \), which is a collection of \( N \) data that belong to a specific domain \( D = \{X, P(x)\} \) with a specific task \( T = \{Y, f(x)\} \).

Often \( P(x) \) and \( f(x) \) are unknown and need to be estimated and learned respectively. If for each sample in the dataset \( N \) its label \( y \in Y \) is given, \( S \) is labelled. Otherwise, \( S \) is unlabelled.
Definition 2.4. (Transfer Learning [170]) "In general, given a source domain $D_S$ and learning task $T_S$, a target domain $D_T$ and learning task $T_T$, transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in $D_T$ using the knowledge in $D_S$ and $T_S$, where $D_S \neq D_T$, or $T_S \neq T_T$. Note that a special topic where $T_S = T_T$ and $D_S$, $D_T$ is known as Domain Adaptation. Specifically, in the context of cross-dataset recognition, the aim of transfer learning is to learn a robust classifier $f(x)$ from a dataset (i.e. target dataset $S_T$) by effectively utilising the knowledge offered through other datasets (i.e. source datasets $S_S$).

2.2 Problem-oriented Taxonomy of Cross-dataset Recognition

In cross-dataset recognition, there are often two datasets. One, referred to as a source dataset, is used in training and the other, referred to as a target dataset, is to be recognized. Their domains and/or tasks are different and their characteristics determines what methods can or should be used. In this paper, we define a set of attributes to characterise the source or target datasets. These attributes have led to a comprehensive taxonomy of cross-dataset recognition problems that provides a unique perspective for this survey.

- **Attributes on data:**
  - **Feature space:** the consistency of feature spaces (i.e. different feature extraction methods or different data modalities) between the source and target datasets.
  - **Data availability:** the availability and sufficiency of target data in the training stage.
  - **Balanced data:** whether the numbers of data samples in each class are balanced.
  - **Sequential/Online data:** whether the data are sequential/online and evolving over time.

- **Attributes on label:**
  - **Label availability:** the availability of labels in source and target datasets.
  - **Label space:** whether the data categories of the two datasets are identical.

Based on these attributes, the following four scenarios are defined as the first layer of the problem taxonomy to guide the survey.

- **Homogeneous feature spaces and label spaces:** The feature spaces and label spaces of the source and target datasets are identical. But domain divergence (i.e. different data distributions) exists across the source and target datasets.
- **Heterogeneous feature spaces:** the feature spaces of the source and target datasets are different (i.e. domain divergence occurs), but their label spaces are the same.
- **Heterogeneous label spaces:** the label spaces of the source and target datasets are different (i.e. task divergence occurs), but their feature spaces are the same.
- **Heterogeneous feature spaces and label spaces:** both the feature spaces and the label spaces of the source and target datasets are different (i.e. both domain and task divergence occurs).

The problems corresponding to the four scenarios are further divided into sub-problems using other data attributes such as the data being balanced and/or sequential/online. Fig. 1 shows the problem-oriented taxonomy for cross-dataset recognition, which shows seventeen different problems.

2.3 Approaches

Many approaches have been developed for transfer learning across datasets [170] at instance level, i.e. re-weighting some source samples based on their divergence from the target domain, at the feature level, i.e. learning “good” feature representations that have minimum domain shift, and at the classifier level, i.e. learn an optimal target classification
Fig. 1. A problem-oriented taxonomy for cross-dataset recognition including the number of papers that are found to address the problems.
model by using the data from both source and target domains as well as the source model. This section summarises several most typical approaches to transfer learning for cross-dataset recognition, including Statistical approach, Geometric approach, Higher-level Representation, Correspondence approach, Class-based approach, Self Labelling, and Hybrid approach. These approaches have been reported explicitly or implicitly in the literature. In particular, the basic assumptions of each approach are analysed and presented in this section. Moreover, several commonly used methods are illustrated under each approach. Due to page limit, only brief description of each approach and its methods is presented. See the supplementary material for details.

**Statistical Approach:** is employed in transferring the knowledge at the levels of instances, features and classifiers by measuring and minimizing the divergence of statistical distributions between the source and target datasets. This approach generally assumes sufficient data in each dataset to approximate the respective statistical distributions. The typical methods are Instance re-weighting [97], Feature space mapping [169] and Classifier parameter mapping [183].

**Geometric Approach:** bridges datasets according to their geometrical properties. It assumes domain shift can be reduced using the relationship of geometric structures between the source and target datasets. Typical methods include Subspace alignment [62], Intermediate subspaces [81, 85], and Manifold alignment (without correspondence) [39].

**Higher-level Representation Approach:** aims at finding higher-level representations that are representative, compact, and invariant between datasets. This approach does not require any labelled data, or the existence of correspondence set, but assumes that there exist the domain invariant higher-level representations between datasets. Note that this approach is commonly used together with other approaches for better transfer, but it is also used independently without any mechanism to reduce the domain divergence explicitly. Typical methods are Sparse coding [185], Low-rank representation [196], Deep Neural Networks [50, 189, 269], Stacked Denoising Auto-encoders (SDAs) [27, 77], and Attribute space [2, 124].

**Correspondence Approach:** uses paired correspondence samples from different domains to construct the relationship between domains. A set of corresponding samples (i.e. the same object captured from different view angles, or by different sensors) are required. The typical methods are Sparse coding with correspondence [285] and Manifold alignment (with correspondence) [271].

**Class-based Approach:** uses label information as a guidance for connecting the source and target datasets. Hence, the labelled data from each dataset are assumed to be available, whether sufficient or not. The commonly used methods include Feature augmentation [44], Metric learning [193], Linear Discriminative Model [264], and Bayesian Model [60].

**Self Labelling:** uses the source domain samples to train an initial model to obtain the pseudo labels of target domain data. Then the target data and their pseudo labels are incorporated to retrain the model. The procedure continues iteratively until convergence. A typical example is Self-training [43, 216].

**Hybrid Approach:** combines two or more above approaches for better transferring of knowledge. Several example combinations are Correspondence and Higher-level representation [96], Higher-level representation and Statistic [147, 148, 243], Statistic and Geometric [273], Statistic and Self labelling [42], Correspondence and Class-based [46], Statistic and Class-based [52], and Higher-level representation and Class-based [288].

In the following sections, we present a comprehensive review on what approaches have been or can be used for the cross-dataset recognition problems shown in Figure 1.
3 HOMOGENEOUS FEATURE SPACES AND LABEL SPACES

In this scenario, \( X_S = X_T \) and \( Y_S = Y_T \). Hence, the \( S \) and \( T \) are generally different in their distributions \( P(X, Y) \). Sufficiently labelled source domain data are generally assumed available and different assumptions are made on the target domain, leading to different sub-problems.

3.1 Labelled Target Dataset

In this problem, a small number of labelled data in target domain are available. However, the labelled target data are generally insufficient for learning an effective classifier. This is also called supervised domain adaptation or few-shot domain adaptation in the literature.

Class-based Approach. The most commonly used approach in supervised domain adaptation is class-based since the labelled data from both domains are available in the training stage. For example, Daumé III [44] propose a feature augmentation based method where each feature is replicated into a high-dimensional space \( \Phi \) containing the general and domain-specific version.

\[
\Phi_s(x) = [x_s, x_s, 0]; \quad \Phi_t(x) = [x_t, 0, x_t];
\]

where \( x_s \in \mathbb{R}^{f \times n_s} \) is the source domain data, \( x_t \in \mathbb{R}^{f \times n_t} \) is the target domain data, \( f \) is the feature dimension, \( n_s \) and \( n_t \) are the total number of samples in the source and target domains, respectively.

The idea of supervised metric learning has also been used [179, 279]. The core idea is to exploit the task relationships between domains to boost the target task. Another group of methods [105, 254, 264] transfer the parameters of discriminative classifiers (e.g. SVM) across datasets. Recently, Motiian et al. [161] propose to create pairs of source and target instances to handle the scarce target labelled data. In addition, they extend adversarial learning [84] to align the semantic information of classes.

A more realistic setting is that samples from only a subset of classes are available in the target domain. Then the adapted features are generalized to unseen categories in the target dataset. While some categories are not available in the target dataset, we still assume the same label spaces between the two domains. So we discuss these methods under the problem of homogeneous label spaces. Generally, these methods assume the shift between domains is category-independent. For example, Saenko et al. [193] present a supervised metric learning-based method to learn a metric that minimizes the distribution shift by using target labelled data from a subset of categories:

\[
\min \text{Tr}(W) - \log \det(W) \quad \text{s.t.} \quad d_W(x_s, x_t) < u \quad \text{if} \quad y_s = y_t, \quad d_W(x_s, x_t) > l \quad \text{if} \quad y_s \neq y_t
\]

where \( u, l \in \mathbb{R} \) are the threshold parameters, \( x_s \) and \( x_t \) represent the source domain sample and target domain sample, respectively and \( y_s \) and \( y_t \) represent their corresponding labels, \( d_W = (x_s - x_t)^TW(x_s - x_t) \) is the distance between \( x_s \) and \( x_t \), and \( W \) is the distance matrix that will be learned. Then the transformation is applied to unseen target test data that may come from different categories from the target training data. Similarly, some recent methods learn to recognize unseen target categories (but have been seen in the source domain) under the deep learning frameworks by exploiting the semantic structure either via soft labels (which is the averaged softmax activations over all source samples in each category) [225] or by the Siamese architecture [162]. For example, Figure 2 illustrates the network architecture of the domain and task transfer method proposed by Tzeng et. al. [225], which uses soft labels. In this work [225], the learned source semantic structure is transferred to the target domain by optimizing the network to produce activation distributions that match those learned for source data.
**Self Labelling.** Dai et al. [43] propose TrAdaBoost to extend boosting-based methods by decreasing the weights of the instances that are most dissimilar to the target distribution in order to weaken their impacts.

**Hybrid Approach.** The higher-level representation approach and class-based approach have been used together for better cross-dataset representation. For example, the discriminative dictionary can be learned such that the same class samples from different domains have similar sparse codes. [198, 290]. Except for the discriminative dictionary learning, the label information can also be used for guiding the deep neural networks to reduce domain shift. For example, Koniusz et al. [116] fuse the source and target CNN streams at the classifier level, where the scatters of the two network streams of the same class are aligned while the between-class are separated.

### 3.2 Labelled plus Unlabelled Target Dataset

Compared to the scenario where only limited labelled target data are presented, additional redundant unlabelled target data are also presented in training in this problem (often known as semi-supervised domain adaptation in the literature) to provide additional structural information. This setting is realistic in real-world applications because unlabelled data are easy to obtain.

**Class-based Approach.** Duan et al. [55] extend SVM-based supervised classifier transfer methods with unlabelled target data. They proposed a regularizer which enforces that the learned target classifiers and the pre-learned source classifiers should have the similar decision values on the unlabelled target instances:

$$
\Omega_D(t_u^T) = \frac{1}{2} \sum_{i=n_l+1}^{n_u} \sum_s \gamma_s (f_i^T - f_s^T)^2 = \frac{1}{2} \sum_s \gamma_s \| f_u^T - f_s^T \|^2
$$

where $t_u^T = [f_{n_l+1}^T, \ldots, f_{n_T}^T]$ and $f_u^s = [f_{n_l+1}^s, \ldots, f_{n_T}^s]$ represent the decision values of the unlabelled target samples from the target classifier and the $s$-th auxiliary classifier, $n_l$ and $n_T$ are the number of labelled target samples and the total number of target samples, $\gamma_s$ is the weight for measuring the relevance between the $s$-th source domain and the target domain.

**Self Labelling.** Some researches extend distance-based classifiers, such as the k-Nearest Neighbour [219] and Nearest Class Mean [38] classifiers, to learn the domain invariant metric iteratively. Specifically, Tommasi and Caputo [219] present a method that learns a metric per class based on the NBNN algorithm, by progressively selecting target instances and combining it with a subset of the source data while imposing a large margin separation hyperplanes among classes. Similarly, Csurka et al. [38] extend the NCM classifier to a Domain Specific Class Means (DSCM) classifier and iteratively add high confidence unlabelled target samples to the training set. A co-training-based method is proposed by [26] to facilitate the gradual inclusion of target features and instances in training. This method iteratively learns feature views and a target predictor upon the views.
Hybrid Approach. A group of methods for semi-supervised domain adaptation combines class-based and statistical approach to make use of both labelled and unlabelled target data. The key idea is that the statistical criteria (e.g. MMD metric between source data and unlabelled target data) are used as an additional constraint in discriminative learning methods (e.g. multiple kernel learning (MKL) [52, 54], or least square method [266]).

Yamada et al. [261] generalize the EASYADAPT method [44] to semi-supervised setting. They proposed to project input features into a higher dimensional space as well as estimate weights for the training samples based on the ratio of test and training marginal distributions in that space using unlabelled target samples.

3.3 Unlabelled Target Dataset

In this problem, no labelled target domain data are available but sufficient unlabelled target domain data are observable for transfer learning. This problem is also named unsupervised domain adaptation. The unsupervised domain adaptation has attracted increasing attention nowadays, which is certainly more realistic and challenging.

Statistical Approach. The Maximum Mean Discrepancy (MMD) criterion is commonly used in unsupervised domain adaptation. Generally, the MMD distance between domains is reduced by re-weighting the samples [78, 97, 213], or mapping to another feature space [9, 150, 169, 273], or regularizing the source domain classifier using target domain unlabelled data [149, 183]. For example, Pan et al. [169] proposed to find a domain invariant feature mapping function \( \phi \) such that the marginal distributions between the two domains \( P_s \) and \( P_t \) in the mapped feature space is small when using the MMD criterion:

\[
D_{\text{MMD}}(P_s, P_t) = \| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi(x_i) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi(x_j) \|_F^2
\]

Except for MMD, other statistical criteria, such as Kullback-Leibler divergence [208], Hellinger distance [10], Quadratic divergence [204], and mutual information [200], are also used for comparing two distributions. Sun et al. [210] propose the CORrelation ALignment (CORAL) to minimize distribution divergence by mapping the covariance of data.

Instead of learning a global transformation, Optimal Transport [36] learns a local transformation such that each source datum is mapped to target data and the marginal distribution is preserved.

Rather than assuming single domain in a dataset, some methods assume a dataset may contain several distinctive sub-domains due to the large variations in visual data. For example, Gong et al. [79] automatically discover latent domains from multi-source domains to characterize the inter-domain variations and, hence, to construct discriminative models.

Geometric Approach. Gopalan et al. [85] proposed a Sampling Geodesic Flow (SGF) method by sampling intermediate subspace representations between the source and target generative subspaces. The two generative subspaces are viewed as two points on a manifold. Then they sample the intermediate subspaces on the geodesic flow between the two subspaces. Lastly, all the data are mapped to the concatenation of all the subspaces to obtain the final representation. Figure 3 illustrates the SGF method. Gong et al. [81] extend SGF to a geodesic flow kernel (GFK) method by proposing a kernel method, such that an infinite number of subspaces are integrated to represent the incremental changes. The methods in [85, 86] and [80, 81] open the opportunity for researches to construct intermediate representations to characterize the domain changes. For example, Zhang et al. [283] bridge the source and target domains by inserting virtual views along a virtual path for cross-view recognition. Rather than manipulating on the subspaces, Cui et al. [40] represent source and target domains as covariance matrices and interpolate some intermediate covariance matrices to bridge the two domains. Some methods [165, 253] are proposed to generate several intermediate domains by learning
the domain-adaptive dictionaries between domains. The idea of intermediate domains is also employed in the deep learning framework [32].

![Illustration of the SGF method (Figure used courtesy of [85])](image)

Instead of modelling intermediate domains, some methods align the two domains directly [4, 39, 62, 153]. For instance, Fernando et al. [62] propose to align the source subspace to the target subspace directly by learning a linear transformation function.

**Higher-level Representation.** The low-rank criterion is commonly used to learn the domain invariant representations [47, 102, 197]. Generally, these methods assume that the data from different domains lie in a shared low-rank structure.

Bengio [14] argue that more transferable features can be learned by deep networks since they are able to extract the unknown factors of variation that are intrinsic to the data. Donahue et al. [50] propose the deep convolutional representations named DeCAF, where a deep CNN model is pre-trained using the source dataset (generally large-scale) in a fully supervised fashion. Then they transfer the features (defined by the pre-learned source convolutional network weights) to the target data. The deep auto-encoders are also used for the cross-dataset tasks by exploiting more transferable features by reconstruction [27, 75, 77, 104, 109]. For instance, Ghifary et al. [75] propose a Deep Reconstruction-Classification Network (DRCN) to learn a shared deep CNN model for both classification task of the source samples and reconstruction task of the target samples.

**Self Labelling.** Recently, Panareda Busto and Gall [171] propose an open set domain adaptation problem, where only some of the classes are shared between the source and target datasets. The task is to label all the target samples either by one of the classes shared between the two domains or as unknown. We discuss this setting under the homogeneous label space problem because the unknown classes are simply detected as unknown rather than recognized as certain classes. They solve this problem by first assigning some of the target data with the labels of the known classes and then reducing the shift between the shared classes in the source and target datasets by a subspace alignment method (similar to [62]). The two procedures are learned iteratively.

**Hybrid Approach.** Combining different approaches generally trigger better transferring of knowledge. Some methods [96, 285] learn two dictionaries on pairs of correspondence samples and encourage the sparse representation of each sample pair to be similar. Some methods use both geometric and statistical approach [211, 273]. For example, Zhang et al. [273] propose to learn two projections for the source and target domain respectively to reduce the geometrical shift and statistical shift. Differently, Gholami et al. [76] jointly learn a low dimensional subspace and a classifier through a Bayesian learning framework.

Though deep networks can generally learn more transferable features [14, 50], the higher level features computed by the last few layers are usually task-specific and are not transferable to new target tasks [269]. Hence, some recent work imposes statistical approach into the deep learning framework (high-level representation approach) to further reduce domain bias. For instance, the MMD loss is incorporated into the objective of the deep models to reduce the divergence
of marginal distributions [148, 152, 227, 229] (e.g. Figure 4 illustrates the Deep Adaptation Networks (DAN) proposed in [148]) or joint distributions [151] between domains.

Instead of using MMD metric, Sun and Saenko [212] extend the CORrelation ALIGNment (CORAL) method [210] that aligns the covariance of the source and target data to a deep learning-based method. Zellinger et al. [270] propose the Central Moment Discrepancy (CMD) method, which align the higher order central moments of distributions through order-wise moment differences. Instead of statistical approach, the self-labelling is also used in deep neural network-based method. Saito et al. [194] propose an asymmetric tri-training method, where feature extraction layers are used to drive three classifier sub-networks. The first two networks are used to label unlabelled target samples and the third network is to learn the final adapted classifier to operate on the target domain with the pseudo-labels obtained on the first two networks.

The statistical approaches (e.g. MMD distance [18, 243], and \(\mathcal{H}\) divergence [18]) are also incorporated into deep autoencoders for learning more transferable features.

Motivated by adversarial learning [84], the GAN-based domain adaptation methods are proposed with the key idea that the JS divergence between domains are reduced [17, 70, 71, 226]. For example, the gradient reversal algorithm (ReverseGrad) proposed by Ganin and Lempitsky [70] minimizes the \(\mathcal{H}\)-divergence by considering the domain invariance as a binary classification task and employing a gradient reversing strategy (as shown in Figure 5). Tzeng et al. [226] propose to learn separate feature extraction networks for different domains, and a domain classifier is incorporated such that the embeddings produced by the source or target CNN cannot be distinguished. Bousmalis et al. [17] propose a GAN-based method to adapt the source domain data from the pixel level, such that they are not distinguishable to the target domain data. Differently, Liu and Tuzel [145] propose a Coupled GAN (CoGAN) method that learns a joint distribution by jointly modelling two GANs, where the first one generates the source data while the second generates the target images. Instead of enforcing samples from different domains to be non-discriminant, the CoGAN enforce the layers that decode high-level features to share the weights so as to enforce the assumption that the images from different domains share the same high-level representations but have different low-level representations.
3.4 Imbalanced Unlabelled Target Dataset

This problem assumes the target domain is class imbalanced and only with unlabelled data. Thus, the statistical approach can be used. This problem is quite common in practice and known as prior probability shift, or imbalanced data in classification. For instance, the abnormal activities (e.g. kick, punch, fight, and fall down) are much less frequent than normal activities (e.g. walk, sit, eat, and drink) in the video surveillance but require higher recognition rate.

**Statistical Approach.** In the classification scenario, the prior probability \( P(Y) \) shift was often considered to be a class imbalance problem [100, 276]. Zhang et al. [276] tackle the prior probability shift by re-weighting the source samples using the similar idea as the Kernel Mean Matching method [97]. They also define the situation where both \( P(Y) \) and \( P(X|Y) \) are shifted across datasets and propose a kernel approach to reduce the distribution shift by re-weighting and transforming the source data. It is assumed that the source data are able to be transferred to the target domain by location-scale (LS) transformation (i.e. \( P(X|Y) \) only differs in the location and scale). Instead of assuming that all the features can be transferred to the target domain by LS transformation, Gong et al. [82] propose to learn the conditional invariant components through a linear transformation, and then the source samples are re-weighted to reduce shift of \( P(Y) \) and \( P(Y|X) \) between domains.

Recently, Yan et al. [263] take both the domain shift and class weight bias across domains into account. To take the class prior probability into account, they introduce class-specific weights. Specifically, the domain adaptation is performed by iteratively generating the pseudo-labels to the target samples, learning the source class weights, and tuning the deep CNN model parameters.

3.5 Sequential/Online Labelled Target Data

In practice, the target data can be sequential video streams or continuous evolving data. The distribution of the target data may also change with time. Since the target data are labelled, this problem is named supervised sequential/online domain adaptation.

**Self Labelling.** Xu et al. [255] assume a weak-labelling setting and propose an incremental method for object detection across domains. Specifically, the adaptation model is a weighted ensemble of the source and target classifiers and the ensemble weights are updated with time.

3.6 Sequential/Online Unlabelled Target Data

Similar to the problem in 3.5, the target data are sequential in this problem, however, no labelled target data is available, which is named unsupervised sequential/online domain adaptation and related to but different from concept drift. The concept of drift [67] refers to changes in the conditional distribution \( P(Y|X) \), while the marginal distribution \( P(X) \) stays unchanged, whereas in sequential/online domain adaptation the changes between the two domains are caused by the changes of the input data distribution.

**Geometric Approach.** Hoffman et al. [92] extend the Subspace Alignment method [62] to handle continuous evolving target domain, as shown in Figure 6. Both the subspaces and subspace metrics that align the two subspaces are updated after each new target sample is received. Bitarafan et al. [15] tackle the continuously evolving target domain using the idea of GFK [81] to construct linear transformation. The linear transformation is updated after a new batch of unlabelled target domain data come. Each batch of arrived target data are classified after the transformation and included in the source domain for recognizing the next batch of data.
represented by a lower dimensional subspace \((CMA)\), which does not require knowledge of discrete do-
swe present our approach, Continuous Manifold Adaptation
parameter to choose the appropriate window size. Next,
that the method was origi-
continuous known domains, where the data from each domain is
drawback of the above methods is that they require dis-
observations.

\[ \int \]

\[ \ell \]

\[ MTDA \]

\[ 53, 55, 79, 93, 214, 257 \]

\[ 166, 167, 260 \]

\[ Xu et al. [260] \]

\[ Multi-Task Autoencoder (MTAE) \]

distinguish from multi-source domain adaptation
since MSDA generally requires the access to the target data for adaptation. We will
discuss transfer learning from multiple sources in details in Section 8.3.

\[ 1960 \]

\[ 1970 \]

\[ 1980 \]

\[ 240 \]

\[ 250 \]

\[ 255 \]

\[ 260 \]

\[ Xu et al. [260] \]

\[ 260 \]

\[ 266 \]

\[ 267 \]

\[ 260 \]

\[ 1950 \]

\[ 44 \]

\[ 50 \]

\[ 3.7 \] Unavailable Target Data

This problem is also named \textit{domain generalization} in literature, where the target domain data are not presented for adaptation. Thus, multiple source datasets are generally required to learn the dataset invariant knowledge that can be
generalized to a new dataset. Note that domain generalization is distinguished from multi-source domain adaptation (MSDA)\cite{53, 55, 79, 93, 214, 257} since MSDA generally requires the access to the target data for adaptation. We will
discuss transfer learning from multiple sources in details in Section 8.3.

\[ Higher-level Representation. \]

Most of the existing work tackle this problem by learning domain invariant and compact representation from multiple source domains \cite{16, 58, 73, 74, 111, 132, 162, 163, 207}. For example, Khosla et al. \cite{111} explicitly model the bias of each source domain and try to estimate the weights for the unbiased data by removing the source domain biases. Muandet et al. \cite{163} propose the Domain-Invariant Component Analysis (DICA), a kernel-based method, to learn an invariant mapping that reduces the domain shift and preserve discriminative information at the same time. Fang et al. \cite{58} propose an unbiased metric learning approach to learn unbiased metric from multiple biased datasets. Ghifary et al. \cite{74} propose a Multi-Task Autoencoder (MTAE) method. It substitutes artificially induced corruption in standard denoising autoencoder with some specific variations of the objects (e.g. rotation) to form multiple views. Hence, MTAE learns representations that are invariant to multiple related domains.

Ensembling classifiers learned from multiple sources is also used for generalizing to unseen target domain \cite{138, 166, 167, 260}. Xu et al. \cite{260} propose to reduce the domain shift in an exemplar-SVMs framework by regularizing positive samples from the same latent domain to have similar likelihoods from each exemplar classifier. Similarly, Niu et al. \cite{166} extend this idea to the source domain samples with multi-view features. Niu et al. \cite{167} explicitly discover the multiple hidden domains \cite{79}, and then an ensemble of classifiers is formed by learning a single classifier for each individual category in each discovered hidden domain.

\[ 4 \text{ HETEROGENEOUS FEATURE SPACES} \]

This section discusses the problems that \(S_S \) and \(S_T \) are different due to \(X_S \neq X_T \), but \(Y_S = Y_T \). The different feature spaces can be generated from different data modalities or different feature extraction methods. Similar to the scenario defined in Section 3, sufficient labelled source domain data are assumed to be available in the following sub-problems.
4.1 Labelled Target Dataset

This problem assumes limited labelled target data are presented for adaptation. This problem is named supervised heterogeneous domain adaptation.

Higher-level Representation. Some methods assume that only the feature spaces are different while the distributions are the same between source and target datasets. Since the labelled data in the target dataset are scarce, Zhu et al. [291] propose to use the auxiliary heterogeneous data that contain both modalities from Web to extract the semantic concept and find the shared latent semantic feature space between different modalities.

Class-based Approach. The class-based approach has also been used to connect heterogeneous feature spaces. Finding the relationship between different feature spaces can be seen as translating between different languages. Hence, Dai et al. [41] propose a translator using a language model to translate between different data modalities or feature spaces by borrowing the class label information. Kan et al. [110] propose a multi-view discriminant analysis method that learns view-specific linear mappings for each view to find a view-invariant space by using label information: 

\[ (w^1, w^2, \ldots, w^n) = \arg \max_w \frac{\text{Tr}(S_B)}{\text{Tr}(S_W)} \]

where the between-class variation \( S_B \) from all views are maximized while the within-class variation \( S_W \) from all views are minimized. \( w^1, w^2, \ldots, w^n \) are the optimized transformations for different views. Manifold alignment method [233] is also used for heterogeneous domain adaptation with the class-based approach.

Inspired by [44], the feature augmentation based method has also been proposed [56, 135] for heterogeneous domain adaptation, which transforms the data from two domains into a shared subspace, and then two transformations are proposed such that the transformed features in the subspace are augmented with the original data as well as zeros (as shown in Figure 7).

\[ \varphi_i(x^s) = \begin{cases} \frac{P(x^s)}{\theta_{s_i}} & \text{if } 0 \leq \| x^s \| \leq \theta_{s_i} \\ 0 & \text{otherwise} \end{cases} \]

\[ \varphi_i(x^t) = \begin{cases} \frac{Q(x^t)}{\theta_{t_i}} & \text{if } 0 \leq \| x^t \| \leq \theta_{t_i} \\ 0 & \text{otherwise} \end{cases} \]

Kulis et al. [118] extend [193] to learn an asymmetric mapping that transforms samples between domains using labelled data from both domains, with the similar assumption as [193] that the label spaces of target training set and target test set are non-overlapping subsets of source label space. Different from previous metric learning based domain adaptation that learns the asymmetric feature transformation between heterogeneous features [118], the asymmetric metric of classifiers can also be learned to bridge source and target classifiers on heterogeneous features [287].

Hybrid Approach. The first group of work focuses on cross-modal representation learning by combing class-based and higher level representation approaches. Gong et al. [83] propose a three-view Canonical Correlation Analysis (CCA) model that explicitly incorporates the high-level semantic information (i.e. high-level labels or topics) as a third view. A recent work [232] incorporates the adversarial learning to the supervised representation learning for cross-modal retrieval.

Another line of research assumes that both the feature spaces and the data distributions are different. Shekhar et al. [199] extend [198] to heterogeneous feature spaces, where the two projections and a latent dictionary are jointly learned to simultaneously find a common discriminative low-dimensional space and reduce the distribution shift. Similarly, Sukhija et al. [209] assume the label distributions low-dimensional space are shared. Then the shared label distributions are used as pivots to derive a sparse projection between the two domains.
4.2 Labelled plus Unlabelled Target Dataset

In this problem, both limited labelled and sufficient unlabelled target data are presented, which is named semi-supervised heterogeneous domain adaptation.

**Statistical Approach.** Tsai et al. [224] propose the Cross-Domain Landmark Selection (CDLS) method for heterogeneous domain adaptation (HDA) using the statistical approach (MMD). Specifically, the CDLS method derives a heterogeneous feature transformation which results in a domain-invariant subspace for associating the heterogeneous domains, and assigns the weight to each instance according to their adaptation ability using both labelled and unlabelled target samples.

**Correspondence Approach.** Zhai et al. [271] assume in addition to a set of labelled correspondence pairs between the source and target datasets, some unlabelled data from both datasets are also available. Specifically, given a set of correspondence samples $C$ between the two domains, one can learn the mapping matrices $A_s$ and $A_t$ for the source and target sets respectively in order to preserve the correspondence relationships after mapping:

$$< A_s, A_t > = \arg \min_{A_s, A_t} \sum_{(i,j) \in C} \| A_s^T x_s^i - A_t^T x_t^j \|^2 + J(A_s, X_s) + J(A_t, X_t)$$

where $x_s^i$ and $x_t^j$ represent the $i$th source domain sample and the $j$th target domain sample, respectively, $J(A_s, X_s)$ and $J(A_t, X_t)$ are the manifold regularization terms which are used to preserve the intrinsic manifold structures of the source and target domains.

**Class-based Approach.** Xiao and Guo [251] propose a kernel matching method, where a kernel matrix of the target domain is matched to a source domain sub-matrix by exploiting the label information such that the target samples are mapped to similar source samples. The unlabelled target samples are expected to be aligned with the source samples from the same class with the guides of labelled target samples via the function of kernel affinity measures between samples.

**Hybrid Approach.** Wu and Ji [247] introduce a constrained deep transfer feature learning method by incorporating the correspondence into the high-level representation approach. Specifically, several pairs of source and target samples are used to capture the joint distribution and bridge the two domains. Then a large amount of additional source samples are transferred to the target domain through pseudo labelling for further target domain feature learning.

4.3 Unlabelled Target Dataset

This problem assumes no labelled target domain data is available. We name this problem as unsupervised heterogeneous domain adaptation. In this problem, the feature spaces could be completely different between datasets. It can also be assumed that the source data consist of multiple modalities while the target data only contain one of the modalities, or vice versa.

**Statistical Approach.** Chen et al. [25] and Li et al. [134] assume the source datasets contain multiple modalities and target dataset only contains one modality and the distribution shift between datasets also exists. Specifically, the statistical approach (e.g. MMD) is used such that the source and target common modalities are projected to a shared subspace to reduce the distribution mismatch. In the meantime, the multiple source modalities are also transformed to the same representation in the shared space. They iteratively refine the shared space and the robust classifier.

**Correspondence Approach.** The co-occurrence data between different feature spaces or modalities have been employed for heterogeneous domain adaptation [180, 265].
Hybrid Approach. The correspondence approach or statistical approach are generally incorporated into higher-level representation approach for transferring between data modalities or feature spaces.

Canonical Correlation Analysis (CCA)[5] is a standard approach to learning two linear projections of two sets of data that are maximally correlated. Neither supervised data nor the paired data are required. Many cross-modal recognition or retrieval methods incorporate the idea of CCA[6, 61, 262] into deep models. Cross-media multiple deep networks (CMDN)[176] jointly preserve the intra-media and inter-media information and then hierarchically combine them for learning the rich cross-media correlation. Castrejón et al. [23] introduce a cross-modal representation method across RGB modality, sketch modality, clipart, and textual descriptions of indoor scenes. The cross-modal convolutional neural networks are regularized using statistical regularization so that they have a shared representation that is invariant to different modalities.

The paired correspondence data are used in [89], where a cross-modal supervision transfer method is proposed. The deep CNNs are pre-trained on the source data (e.g. a large-scale labelled RGB dataset). Then the paired target data (unlabelled RGB and depth image pairs) are used for transferring the source parameters to the target networks by constraining the paired samples from different modalities to have the similar representations.

A line of research focuses on the task of translation between different domains. For example, in machine translation between languages, the sentence pairs are presented in the form of a parallel training corpus for learning the translation system. Traditional translation system [115] is generally phrase-based, whose sub-components are usually learned separately. Differently, a newly emerging approach, named Neural machine translation [8, 31, 108, 215], constructs and trains a neural network that inputs a sentence and outputs the translated sentence.

Similarly, in the computer vision domain, image-to-image translation [98] has also been extensively exploited, which aims at converting an image from one representation of a given scene to another (e.g. texture synthesis [130], sketch to photograph [98], RGB to depth [89], time hallucination [98, 122, 201], image to semantic labels [57, 146, 252], stimulated to real image [203], style transfer [72, 107, 130, 239, 278], and general image-to-image translation [13, 98, 112, 129, 144, 145, 268, 289]). The key idea for tackling these tasks is to learn a translation model between paired (correspondence approach) or unpaired samples (statistical approach) from different domains. The recent deep learning based techniques have greatly advanced the image-to-image translation task. For example, the deep convolutional neural networks based methods [57, 72, 107, 146, 252, 278], and the Generative Adversarial Networks (GANs [84]) based methods [13, 98, 112, 129, 130, 144, 145, 203, 239, 268, 289] have been exploited for learning the translation model. Though the original purposes of some of these work on translation between domains may not be cross-dataset recognition, the ideas can be borrowed for cross-modality or cross feature spaces recognition. If a proper translation between domains can be obtained, the target task can be boosted by the translated source domain data.

5 HETEROGENEOUS LABEL SPACES

This section discusses the problems that $X_S = X_T$ and $Y_S \neq Y_T$. For example, in the classification tasks, when the label spaces between datasets are different, there still exists shared knowledge between previous categories (e.g. horse) and new categories (e.g. zebra) that can be used for learning new categories. The source domain is assumed to be labelled except for the last sub-problem (Section 5.5).

5.1 Labelled Target Dataset

This setting is commonly used in deep learning context. In practice, the deep networks are rarely trained from scratch (with random initialization), since the target datasets rarely have sufficient labelled data. Thus, transfer learning is
generally used. The pre-trained deep models from a very large source dataset are used either as an initialization (then fine-tune the model according to the target data) or as a fixed feature extractor for the target, which is generally different from the original task (i.e., different label spaces).

The fine-tuning procedure is similar to one-shot learning or few-shot learning. The key difference is that the available target data are sufficient for the target task in fine-tuning but in few-shot learning, the target data are generally rare (e.g., only one sample per class in the extreme case). The few-shot learning also has close connection with multi-task learning. The difference is that one-shot learning emphasizes on the recognition of the target data with limited labelled data while the objective of multi-task learning is to improve all the tasks with good training data in each task.

Higher-level Representation Approach. Since the training of deep learning models requires a large scale dataset to avoid overfitting, the transfer learning techniques [269] can be used for small scale target datasets. The most commonly used transfer learning technique is to initialize the weights from a pre-trained model and then the target training data are used to fine-tune the parameters for the target task. When the pre-trained source model is used as the initialization, two strategies can be employed. First is to fine-tune all the layers of the deep neural network, while the second strategy is to freeze several earlier layers and only fine-tune the later layers to reduce the effects of overfitting. This is inspired by the observation that the features extracted from the early layers show more general features (e.g., edge or color) that are transferable to different tasks. However, the later layers are gradually more specific to the details of the original source tasks. Other transfer methods [50, 189] directly use the pre-trained deep convolutional nets (normally after removing the last one or two fully connected layers) on a large dataset (e.g., ImageNet [45]) as a fixed feature extractor for the target data.

Note that when the pre-trained deep models are used as an initialization or a fixed feature extractor in the deep learning frameworks, only the pre-trained weights need to be stored without the need of storing the original large scale source data, which is appealing.

Class-based Approach. Patricia and Caputo [174] treat the pre-trained models from multi-source domains as experts to augment the target features. The output confidence values of prior models are treated as features and the features from the target samples are augmented with these confidence values to build a target classifier. Several classifier-based methods are proposed to transfer the parameters of classifiers using generative models [60, 123], or discriminative models [7, 106, 156, 220]. The key idea is using source models as prior knowledge to regularize the models of the target task. These methods are also called the Hypothesis Transfer Learning (HTL) since it assumes no explicit access to the source domain data and only uses source models learned from a source domain. The HTL has been theoretically analysed [51, 121, 241].

Hybrid Approach. Recently, the deep learning based approaches have been proposed for few-shot learning, most of which are metric learning based methods. One early neural network approach to one-shot learning was provided by Siamese networks [113], which employs a structure to rank similarity between inputs. Vinyals et al. [230] propose the matching networks, where a differentiable neural attention mechanism is used over a learned embedding of the limited labelled target data. This method can be considered as a weighted nearest-neighbour classifier in an embedded space. Snell et al. [205] transform the input into an embedding space by proposing a prototypical network and the prototype from each class is taken as the mean of the embedded support set. Differently, Ravi and Larochelle [188] propose a meta-learning-based few-shot learning method, where a meta-learner LSTM [91] model is used to produce updates for training the few-shot neural network classifier. Given a few target labelled examples, this approach can generalize well on the target set.
5.2 Unlabelled Target Dataset

Some researches also try to tackle the heterogeneous label space problem by assuming that only unlabelled target data are presented. This problem can be named as unsupervised transfer learning.

Higher-level Representation. The higher-level representation approach is generally used for this problem. Two different scenarios are considered in literature.

The first scenario assumes that only the label spaces between datasets are disjoint while the distribution shift is not considered. Since no labelled target data are available, the unseen class information is generally gained from a higher level semantic space shared between datasets. For example, some research assumes that the human-specified high-level semantic space (e.g. attributes [168], or text descriptions [190]) shared between datasets are available. Given a defined attribute or text description ontology, a vector in the semantic space can be used for representing each class. However, it is expensive to acquire the attribute annotations or text descriptions. Hence, to avoid human involved annotations, another strategy learns the semantic space by borrowing the large and unrestricted, but freely available, text corpora (e.g. Wikipedia) to derive a word vector space [64, 158, 206]. The related work on semantic space (e.g. attributes, text descriptions, or word vector) will be further discussed in Section 5.4, since the target data are generally not required when the semantic space is involved.

The second scenario assumes that apart from the different label spaces, the domain shift (i.e. the distribution shift of features) also exists between datasets [66, 114, 139, 234, 259, 267, 282]. This is named the projection domain shift problem by [66]. For example, as illustrated in Figure 8, both zebra and pig have the same attribute ’hasTail’, but the visual appearances and the distributions of the tails of zebra and pig are very different. To reduce the domain shift explicitly, the training data (unlabelled) in the target domain are generally required to be available. For example, Fu et al. [66] introduce a multi-view embedding space in a transductive setting, such that different semantic views are aligned. Kodirov et al. [114] propose a regularised sparse representation framework that utilizes the target class prototypes estimated from target images to regularise the projections of the target data and thus overcomes the projection domain shift problem.

![Fig. 8. Examples of projection domain shift (Figure used courtesy of [66])](image)

5.3 Sequential/Online Labelled Target Data

This problem assumes the target data are sequential and can be from different classes, which is also called sequential/online transfer learning, and closely related to lifelong learning [141, 192, 218]. Both concepts focus on the continuous learning processes for evolving tasks. However, sequential/online transfer learning emphasizes on how to improve the target domain performance (without sufficient target training data), but lifelong learning tries to improve the future target
Self Labelling. Nater et al. [164] address an action recognition scenario where the unseen activities to be recognized only have one labelled sample per new activity. They build a multi-class model which uses the prior knowledge of seen classes and progressively learns the new classes. Then the newly labelled activities are integrated into the previous model to update the activity model. Zhao and Hoi [284] propose an ensemble learning based online transfer learning method (OTL) that learns a classifier in an online fashion using the target data, and combines it with the pre-learned source classifier. The combination weights are tuned dynamically based on the loss between the ground-truth label of the incoming sample and the current prediction. Tommasi et al. [221] then extended OTL [284] and addressed the case of online transfer learning from multiple sources.

5.4 Unavailable Target Data

This problem is also named zero-shot learning in literature, where unseen target categories are to be recognized without having access to the target data. Different from domain generalization (see Section 3.7), the categories of unseen target data are different from the source categories in zero-shot learning. As mentioned in Section 5.2, the unseen categories can be generally connected via some auxiliary information, such as a common semantic space.

Higher-level Representation. Most of the methods for this problem rely on the existence of a labelled source dataset of seen categories and the prior knowledge about the semantic relationship between the unseen and seen categories. In general, the seen and unseen categories are correlated in a high-level semantic space. Such a semantic space can be an attribute space [168], text description space [190], or a word vector space [64, 158, 206]. Since multiple semantic spaces are often complementary to each other, some methods are proposed to fuse multiple semantic spaces [3, 277].

The attribute space is the most commonly used intermediate semantic space. The attributes are defined as properties observable in images, which are described with human-designated names such as "white", "hairy", "four-legged". Hence, in addition to label annotation, the attribute annotations are required for each class. However, the attributes are annotated per-class rather than per-image. Thus, the effort to annotate a new category is small. Two main strategies are proposed for recognizing unseen object categories using attributes. The first is recognition using independent attributes, consists of learning an independent classifier per attribute [120, 124, 143, 168, 172]. At test time, the attribute values for test data are predicted using the independent classifiers and the labels are then inferred. Since attribute detectors are expected to generalize well on both seen and unseen categories, some research is devoted to discovering discriminant attributes [24, 182, 187], or modelling the uncertainty of attributes [101, 240], or robustly detecting attributes from images [19, 68]. However, Akata et al. [2] argue that the attribute classifiers in previous works are learned independently of the end-task, and thus they may be able to predict the attributes from new images but may not be able to effectively infer the classes. Hence, the second strategy is recognition by assuming a fixed transformation (W) between the attributes and the class labels [1, 3, 140, 181, 191, 248, 280, 281] to learn all attributes simultaneously: \( F(x, y; W) = \theta(x)^T W \phi(y) \), where \( \theta(x) \) and \( \phi(y) \) represent image and class embeddings, both are given. To sum up, the attribute-based zero-shot learning methods are promising for recognizing unseen classes, while with a key drawback that the attribute annotations are still required for each class. Instead of using attributes, the second semantic space is image text descriptions [190], which provides a natural language interface. However, similar to attribute space, the expensive manual annotation is required for obtaining the good performance. The third semantic space is the word vector space [64, 128, 158, 206], which is derived from a huge text corpus and generally learned by a deep neural network. The word vector space is attractive since extensive annotations are not required for obtaining the semantic space.
5.5 Unlabelled Source Dataset

This problem assumes that the source data are unlabelled but the contained information (e.g. basic visual patterns) can be used for target tasks, which is known as self-taught learning.

*Higher-level Representation.* Raina et al. [185] firstly presented the idea of “self-taught learning”. They learn the sparse coding from the source data to extract higher-level features. Some variations of Raina et al. [185]’s method are proposed either by generalizing the Gaussian sparse coding to exponential family sparse coding [126], or by taking the supervision information contained in labelled images into consideration [235]. Moreover, Kumagai [119] provide a theoretical analysis for self-taught learning with the focus on discussing the learning bound of sparsity-based methods.

The idea of self-taught learning has also been used in deep learning framework, where the unlabelled data are used for pre-training the network to obtain good starting point of parameters [69, 117, 125]. For instance, Gan et al. [69] use the unlabelled samples to pre-train the first layer of Convolutional deep belief network (CDBN) for initializing the network parameters. Kuen et al. [117] extract the domain-invariant features from unlabelled source image patches for the tracking tasks using stacked convolutional autoencoders.

6 HETEROGENEOUS FEATURE SPACES AND LABEL SPACES

In this section, a more challenging scenario is discussed, where \( X_S \neq X_T \) and \( Y_S \neq Y_T \). There is little work regarding this scenario due to the challenges and the common assumption that sufficient source domain labelled data is available.

6.1 Labelled Target Dataset

This problem assumes the labelled target data are available. We name this problem as heterogeneous supervised transfer learning.

*Higher-level Representation.* Rather than assuming completely different feature spaces, most methods in this setting assume that the source domain contains data with multi-modality but the target domain only has one of the source domain modalities. Ding et al. [48] propose to uncover the missing target modality by finding similar data from the source domain, where a latent factor is incorporated to uncover the missing modality based on the low-rank criterion (as illustrated in Figure 9). Similarly, Jia et al. [103] propose to transfer the knowledge of RGB-D (RGB and depth) data to the dataset that only has RGB data. They applied the latent low-rank tensor method to discover the common subspace of the two datasets.

*Hybrid Approach.* Hu and Yang [95] assume the feature spaces, the label spaces, as well as the underlying distributions are all different between source and target datasets and propose to transfer the knowledge between different activity recognition tasks by learning a mapping between different sensors. They adopt the similar idea of translated learning [41] to find a translator between different feature spaces using statistical approach (e.g. JS divergence). Then the Web knowledge is used to link the different label spaces using self-labelling.
6.2 Sequential/Online Labelled Target Data

This problem assumes the sequential/online target data have different feature space with source data, which is named as heterogeneous sequential/online transfer learning.

Self Labelling. As mentioned in Section 5.3, Zhao and Hoi [284] propose the OTL method for online transfer learning. They also consider the case of heterogeneous feature spaces by assuming the source domain feature space to be a subspace of the target domain feature space. Then a multi-view approach is proposed by adopting a co-regularization principle of online learning of two target classifiers simultaneously from the two views (the source domain feature space and the new space). The unseen target example is classified by the combination of the two target classifiers.

7 DATASETS

Table 1 lists the commonly used visual datasets for transfer learning. They are categorised into object recognition, Hand-Written digit recognition, face recognition, person re-identification, scene categorization, action recognition and video event detection. In the table, the ✓ indicates the dataset has been evaluated on the corresponding problem while the # indicates the datasets that have the potential to be used in the evaluation of the algorithms for the problem though reported results are not publicly available to our knowledge. Due to the page limit, readers are referred to the supplementary material and the references for more detailed information of the datasets.

8 CHALLENGES AND FUTURE DIRECTIONS

Transfer learning is a promising and important approach to cross-dataset visual recognition and has been extensively studied in the past decades with much success. Figure 1 shows the problem-oriented taxonomy and the statistics on the number of papers for each problem has showed that most previous works concentrate on a subset of problems presented in Figure 1. Specifically, only nine out of the seventeen problems are relatively well studied where the source and target domains share at least either their feature spaces or label spaces, the source domain data are labelled and balanced, target domain data are balanced and non-sequential. The rest eight problems especially those where the target data is imbalanced and sequential are much less explored. Such a landscape together with the recent fast-advancing deep learning approach has revealed many challenges and opened many future opportunities as elaborated below for cross-dataset visual recognition.

8.1 Deep Transfer Learning

As deep learning advances, transfer learning is also shifted from traditional shallow-learning based approaches to deep neural network based approaches. In practice, the deep networks for the target task are rarely trained from scratch (i.e. with random initialization), since the target datasets rarely have sufficient samples. Thus, transfer learning is generally used. The pre-trained deep models from a very large source dataset are used either as an initialization [269] (then fine-tune the model according to the target data) or a fixed feature extractor for the target task of interest [50, 189].

Similarly, in deep domain adaptation, the deep models are either used as feature extractors (then shallow-based domain adaptation methods are used for further adaptation) [36, 73, 116, 224, 266, 273], or used in an end-to-end fashion (i.e. the domain adaptation module is integrated into the deep model) [17, 71, 148, 151, 152, 225, 226]. It is still unclear which approach would perform better. The advantage of using deep models as feature extractors is that the computational cost is much lower since shallow-based DA methods are generally much faster than deep learning-based methods. Another advantage is that many shallow-based methods have a global optimum value. The drawback is that the degree of adaptation may be insufficient in the shallow-based methods to fully leverage the deeply extracted features. On the other hand, the advantage of integrating an adaptation module into deep models is two-fold. First, it is
end-to-end trainable. Secondly, the adaptation can be performed in multiple levels of features. While the drawbacks are the computational cost and the local optimum. To date, these two approaches have produced similar performance on some datasets [116, 152, 162, 274] though the end-to-end deep systems involve more parameters and require more computational costs. One of the missing study in the literature is a systematic study and comparison of the two approaches under same or similar conditions. For instance, both deep and shallow-based methods can use MMD metric between distributions as a constraint to the objective function. Thus, the comparison between the two approaches using MMD metric may be conducted.

Table 1. Suitability of the widely used datasets where the ✓ indicates the dataset has been used the corresponding problems while the # indicates the datasets can be potentially used for the problem. Problem notations: P3.1, supervised domain adaptation (DA); P3.2, Semi-supervised DA; P3.3, Unsupervised DA; P3.4, Supervised online DA; P3.5, Supervised online DA; P3.6, Unsupervised online DA; P3.7, Domain generalization; P4.1, Supervised Heterogeneous DA; P4.2, Semisupervised Heterogeneous DA; P4.3, Unsupervised Heterogeneous DA; P5.1, Few-shot Learning; P5.2, Unsupervised transfer learning (TL); P5.3, Online TL; P5.4, Zero-shot Learning; P5.5, Self-taught Learning; P6.1, Heterogeneous TL; P6.2, Heterogeneous online TL.
The adversarial nets derived from GANs [84] are appealing in deep learning-based transfer methods. The adversarial loss measures the JS divergence between two sets of data. In practice, the adversarial loss achieves better results and requires smaller batch sizes compared to the MMD loss [71, 131]. Currently, the adversarial nets-based transfer methods have been used on many transfer learning tasks, such as domain adaptation [17, 70, 71, 145, 226], partial domain adaptation [21, 272], cross-modal transfer [13, 98, 112, 129, 130, 144, 145, 203, 239, 268, 289], and zero-shot learning [249, 292]. However, some of the drawbacks of GANs may also remain in adversarial nets-based transfer methods, such as unclear stopping criteria and hard training.

8.2 Partial Domain Adaptation

Partial domain adaptation aims at adapting from a source dataset to an unlabelled target dataset whose label space is known to be a subspace of that of the source [21, 94, 272] or in a more general and challenging setting where only a subset of the label spaces between the source and target is overlapping [171]. The former may be considered to be a special case of transfer learning between heterogeneous label spaces and a typical and practical example is to transfer from a large source dataset with more classes to a small target dataset with less classes. The latter is a problem bearing both domain adaptation and zero shot learning. Generally, the distribution shift is caused not only by label space difference but also by the intrinsic divergence of distributions (i.e., the distribution shifts exist even on shared classes between source and target). Partial domain adaptation has a more realistic setting than conventional unsupervised domain adaptation. Solutions to this problem would expand the applications of domain adaptation and provide a basic mechanism for online transfer learning and adaptation. However, few papers have been found on partial domain adaptation.

8.3 Transfer Learning from Multiple Sources

The multi-source domain adaptation (MSDA) [53, 55, 79, 93, 214, 257] refers to adaptation from multiple source domains that have exactly the same label space as the target domain. Intuitively, the MSDA methods should be able to obtain superior performance compared to the single source setting. However, in practice, the adaptation from multiple sources generally can only give similar or even worse results compared to transferring from one of the source domains (though not every one of them) [102, 198]. This is probably due to the negative transfer issue. In addition, most source data contains multiple unknown latent domains [79, 93] in the real-world applications. Thus, how to discover latent domains and how to measure the domain similarities are still fundamental issues.

A more realistic setting is incomplete multi-source domain adaptation (IMSDA) [49, 257] here each source label space is only a subset in the target domain and the union of the multiple source label spaces covers the target label space. IMSDA is a more challenging problem compared with MSDA, since the distribution shifts among the sources as well as the target domain are harder to be reduced due to the incompleteness of each source domain. In addition, when the number of sources increases, this problem will become challenging.

Multiple sources can be generalised to a target task, referred to as domain generalization [16, 58, 73, 74, 111, 132, 162, 163, 207] without the need of any target data. Domain generalization is of practical significance, but less addressed in the previous research. Since there is no target data available, domain generalization often has to learn semantically meaningful model shared across different domains.

8.4 Sequential/Online Transfer Learning

In sequential/online transfer learning [284], source data may not be fully available when the adaptation or transfer learning is being performed and/or the target data may also arrive sequentially. In addition, the source or even the
target data cannot be fully stored and revisited in the future learning process. The adapted model is often required to perform well not only on the new target data but also to maintain its performance on the source data or previously seen data. Such a setting is sometimes known as incremental learning or transfer learning without forgetting under certain assumptions [127, 141, 202]. Few studies on this problem have been reported as shown in Figure 1.

8.5 Data Imbalance

The issue of data imbalance in the target dataset has been much neglected in the previous research, while imbalanced source data may be converted to balanced ones by discarding or re-weighting the training (source) data during the learning procedure. However, the target data can hardly follow such a process especially when the target data is insufficient. Data imbalance can be another source of distribution divergence between datasets and is ubiquitous in real-world applications. So far, there has been little study on how the existing algorithms for cross-dataset recognition would perform on imbalanced target data or how the imbalance would affect the algorithm performance.

8.6 Few-shot and Zero-shot Learning

Few-shot learning and Zero-shot learning are interesting and practical sub-problems in transfer learning which aim to transfer the source models efficiently to the target task with only a few (few-shot) or even no target data (zero-shot). In few-shot learning, the target data are generally rare (i.e. only one training sample is available for each class in the extreme case). Thus, the standard supervised learning framework could not provide an effective solution for learning new classes from only few samples [60, 123]. This challenge becomes more obvious in the deep learning context, since it generally relies on larger datasets and suffers from overfitting with insufficient data [205, 230].

Compared to few-shot learning, zero-shot learning does not require any target data. A key challenge in zero-shot learning is the issue of projection domain shift [65], which is neglected by most previous work. Since the source and target categories are disjoint, the projection obtained from the source categories is biased if they are applied to the target categories directly. For example, both zebra (one of the source class) and pig (one of the target class) have the same attribute ‘hasTail’, but the visual appearances of the tails of zebra and pig are very different (as shown in Figure 8). However, to deal with the projection domain shift problem, the unlabelled target data are generally required. Thus, further exploration of new solutions to reduce the projection domain shift is useful for effective zero-shot learning.

Another future direction is the exploration of more high-level semantic spaces for connecting seen and unseen classes. The most frequently used high-level semantics are manually annotated attributes or text descriptions. Some recent work [64, 128, 158, 206] employs the word vector as semantic space without relying on human annotation, but the performance of zero-shot learning using word vector is generally poorer than that using manually labelled attributes.

A recent work [250] presents a comprehensive analysis of the recent advances in zero-shot learning. They critically compare and analyse the state-of-the-art methods and unifies the data splits of training and test sets as well as the evaluation protocols for zero-shot learning. Their evaluation protocol emphasizes on the generalized zero-shot learning, which is considered more realistic and challenging. The traditional zero-shot learning generally assumes that the training categories do not appear at test time. By contrast, the generalized zero-shot setting relaxes this assumption and generalizes to the case where both seen and unseen categories are presented in the test stage, which provides standard evaluation protocols and data splits for fair comparison and realistic evaluation in the future.

8.7 Cross-modal Recognition

The cross-modal transfer, a sub-problem of heterogeneous domain adaptation and heterogeneous transfer learning as shown in Figure 1, refers to transfer between different data modalities (e.g. text v.s. image, image v.s. video, RGB
v.s. Depth, etc.). Compared to cross-modal retrieval [232] and translation [98], fewer works are dedicated to cross-modal recognition through adaptation or transfer learning. The recognition across data modalities is ubiquitous in the real-world applications. For instance, the depth images acquired by the newly released depth cameras are much rarer compared to RGB images. Effectively using rich and massive labelled RGB images to help the recognition of depth images can reduce the extensive efforts of data collection and annotation. Some preliminary works can be found in [89, 103, 129, 134, 238].

8.8 Transfer Learning from Weakly Labelled Web Data
The data on the Internet are generally weakly labelled. Textual information (e.g., caption, user tags, or description) can be easily obtained from the web as additional meta information for visual data. Thus, effectively adapting the visual representations learned from the weakly labelled data (e.g. web data) or co-existent other modality data to new tasks is interesting and practically important. A recent work releases a large scale weakly labelled web image dataset (WebVision [136]).

8.9 Self-taught Learning
A natural assumption among most of the literature is that the source data are extensive and labelled. This may be because the source data are generally treated as the auxiliary data for instructing or teaching the target task and the unlabelled source data could be unrelated and may lead to negative transfer. However, some research works argue that the redundant unlabelled source data can still be a treasure as a good starting point of parameters for target task as mentioned in Section 5.5. How to effectively leverage the massively available unlabelled source data to improve the transfer learning approaches is an interesting problem.

8.10 Large Scale and Versatile Datasets for Transfer Learning
The development of algorithms usually depends very much on the available datasets for evaluation. Most of the current visual datasets for cross-dataset recognition are small scale in terms of either number of classes or number of samples and they are especially not suitable for evaluating deep learning algorithms. An establishment of truly large scale versatile (i.e. suitable for different problems) and realistic dataset would drive the research a significant step forward. As well known, the creation of a large scale dataset may be unaffordably expensive. Combinations and re-targeting of existent datasets can be an effective and economical way as demonstrated in [275]. As shown in Table 1, there are few visual recognition datasets designed for online transfer learning (e.g. P3.5, P3.6, P5.3, and P6.2). Most of the current online transfer learning deals with the detection tasks[255] or text recognition tasks[284]. To advance the transfer learning approaches for more broad and realistic applications, it is essential to create a few large scale datasets for online transfer learning.

9 CONCLUSION
Transfer learning from previous data for current tasks has a wide range of real-world applications. Many transfer learning algorithms for cross-dataset visual recognition have been developed in the last decade as reviewed in this paper. A key question that often puzzles a practitioner or a researcher is that which algorithm should be adopted for a given task. This paper intends to answer the question by providing a problem-oriented taxonomy of transfer learning for cross-dataset recognition and a comprehensive survey of the recently developed algorithms with respect to the taxonomy. Specifically, we believe the choice of an algorithm for a given target task should be guided by the attributes of both source and target datasets and the problem-oriented taxonomy offers an easy way to look up the problem and
the methods that are likely to solve the problem. In addition, the problem-oriented taxonomy has also shown that many challenging problems in transfer learning for visual recognition have not been well studied. It is likely that research will focus on these problems in the future.

Though it is impossible for this survey to cover all the published papers on this topic, the selected works have well represented the recent advances and in-depth analysis of these works have revealed the future research directions in transfer learning for cross-dataset visual recognition.

ACKNOWLEDGMENTS

This work is partially supported by the Australian Research Council Future Fellowship under Grant FT180100116.

REFERENCES

Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective


Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective


Jing Zhang, Wanqing Li, Philip Ogunbona, and Dong Xu


Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective


[204] Si Si, Dacheng Tao, and Bo Geng. 2010. Bregman divergence-based regularization for transfer subspace learning. IEEE Transactions on Knowledge and Data Engineering 22, 7 (2010), 929–942.


Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective


Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective


