

Domain Adaptation for Visual Applications: A Comprehensive Survey

Gabriela Csurka

Abstract *The aim of this paper¹ is to give an overview of domain adaptation and transfer learning with a specific view on visual applications. After a general motivation, we first position domain adaptation in the larger transfer learning problem. Second, we try to address and analyze briefly the state-of-the-art methods for different types of scenarios, first describing the historical shallow methods, addressing both the homogeneous and the heterogeneous domain adaptation methods. Third, we discuss the effect of the success of deep convolutional architectures which led to new type of domain adaptation methods that integrate the adaptation within the deep architecture. Fourth, we overview the methods that go beyond image categorization, such as object detection or image segmentation, video analyses or learning visual attributes. Finally, we conclude the paper with a section where we relate domain adaptation to other machine learning solutions.*

1 Introduction

While huge volumes of unlabeled data are generated and made available in many domains, the cost of acquiring data labels remains high. To overcome the burden of annotation, alternative solutions have been proposed in the literature in order to exploit the unlabeled data (referred to as semi-supervised learning), or data and/or models available in similar domains (referred to as transfer learning). Domain Adaptation (DA) is a particular case of transfer learning (TL) that leverages labeled data in one or more related *source* domains, to learn a classifier for unseen or unlabeled data in a *target* domain. In general it is assumed that the task is the same, *i.e.* class labels are shared between domains. The source domains are assumed to be related to the target domain, but not identical, in which case, it becomes a standard machine learning (ML) problem where we assume that the test data is drawn from the same distribution as the training data. When this assumption is not verified, *i.e.* the distributions of training and test sets do not match, the performance at test time can be significantly degraded.

Xerox Research Center Europe (www.xrce.xerox.com), 6 chemin Maupertuis, 38240 Meylan, France, e-mail: Gabriela.Csurka@xrce.xerox.com

¹ Book chapter to appear in "Domain Adaptation in Computer Vision Applications", Springer Series: Advances in Computer Vision and Pattern Recognition, Edited by Gabriela Csurka.

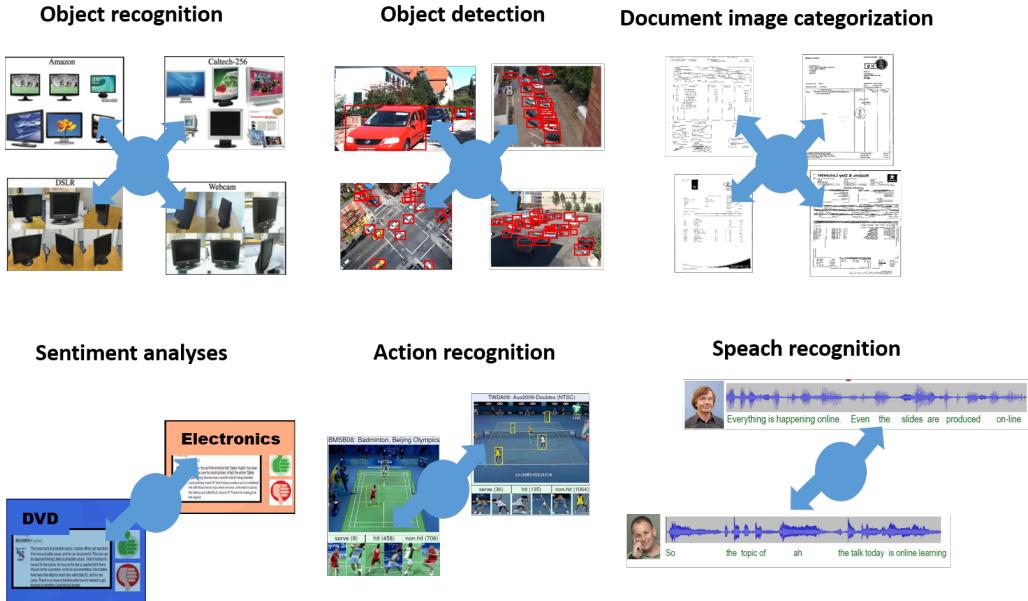


Fig. 1 Example scenarios with domain adaptation needs.

In visual applications, such distribution difference, called domain shift, are common in real-life applications. They can be consequences of changing conditions, *i.e.* background, location, pose changes, but the domain mismatch might be more severe when, for example, the source and target domains contain images of different types, such as photos, NIR images, paintings or sketches [1, 2, 3, 4]. Service provider companies are especially concerned since, for the same service (task), the distribution of the data may vary a lot from one customer to another. In general, machine learning components of service solutions that are re-deployed from a given customer or location to a new customer or location require specific customization to accommodate the new conditions. For example, in brand sentiment management it is critical to tune the models to the way users talk about their experience given the different products. In surveillance and urban traffic understanding, pretrained models on previous locations might need adjustment to the new environment. All these entail either acquisition of annotated data in the new field or the calibration of the pretrained models to achieve the contractual performance in the new situation. However, the former solution, *i.e.* data labeling, is expensive and time consuming due to the significant amount of human effort involved. Therefore, the second option is preferred when possible. This can be achieved either by adapting the pretrained models taking advantage of the unlabeled (and if available labeled) target set or, to build the target model, by exploiting both previously acquired labeled source data and the new unlabeled target data together.

Numerous approaches have been proposed in the last years to address adaptation needs that arise in different application scenarios (see a few examples in Figure 1). Examples include DA and TL solutions for named entity recognition and opinion extraction across different text corpora [5, 6, 7, 8], multilingual text

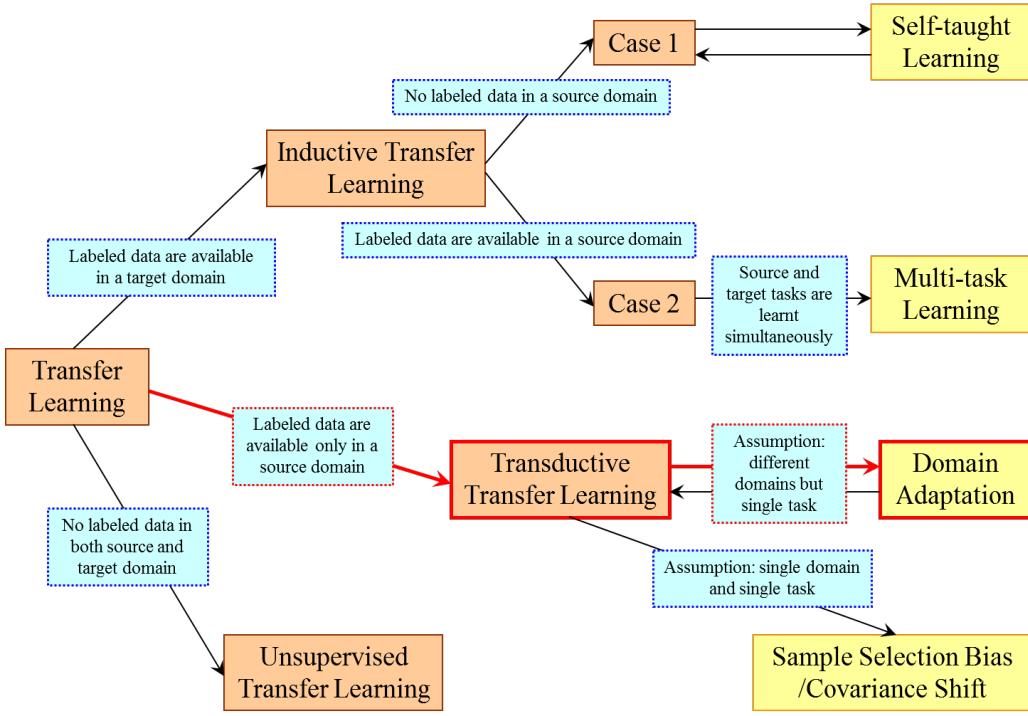


Fig. 2 An overview of different transfer learning approaches. (Image: Courtesy to S.J. Pan [37].)

classification [9, 10, 11], sentiment analysis [12, 13], WiFi-based localization [14], speech recognition across different speakers [15, 16], object recognition in images acquired in different conditions [17, 18, 19, 20, 21], video concept detection [22], video event recognition [23], activity recognition [24, 25], human motion parsing from videos [26], face recognition [27, 28, 29], facial landmark localization [30], facial action unit detection [31], 3D pose estimation [32], document categorization across different customer datasets [33, 34, 35], etc.

In this paper, we mainly focus on *domain adaptation* methods applied to *visual tasks*. For a broader review of the transfer learning literature as well as for approaches specifically designed to solve non-visual tasks, *e.g.* text or speech, please refer to [36].

The rest of the paper is organized as follows. In Section 2 we define more formally transfer learning and domain adaptation. In Section 3 we review shallow DA methods that can be applied on visual features extracted from the images, both in the homogeneous and heterogeneous case. Section 4 addresses more recent deep DA methods and Section 5 describes DA solutions proposed for computer vision applications beyond image classification. In Section 6 we relate DA to other transfer learning and standard machine learning approaches and in Section 7 we conclude the paper.

2 Transfer learning and domain adaptation

In this section, we follow the definitions and notation of [37, 36]. Accordingly, a domain \mathcal{D} is composed of a d -dimensional feature space $\mathcal{X} \subset R^d$ with a marginal probability distribution $P(\mathbf{X})$, and a task \mathcal{T} defined by a label space \mathcal{Y} and the conditional probability distribution $P(\mathbf{Y}|\mathbf{X})$, where \mathbf{X} and \mathbf{Y} are random variables. Given a particular sample set $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of \mathcal{X} , with corresponding labels $\mathbf{Y} = \{y_1, \dots, y_n\}$ from \mathcal{Y} , $P(\mathbf{Y}|\mathbf{X})$ can in general be learned in a supervised manner from these feature-label pairs $\{\mathbf{x}_i, y_i\}$.

Let us assume that we have two domains with their related tasks: a *source* domain $\mathcal{D}^s = \{\mathcal{X}^s, P(\mathbf{X}^s)\}$ with $\mathcal{T}^s = \{\mathcal{Y}^s, P(\mathbf{Y}^s|\mathbf{X}^s)\}$ and a *target* domain $\mathcal{D}^t = \{\mathcal{X}^t, P(\mathbf{X}^t)\}$ with $\mathcal{T}^t = \{\mathcal{Y}^t, P(\mathbf{Y}^t|\mathbf{X}^t)\}$. If the two domains corresponds, *i.e.* $\mathcal{D}^s = \mathcal{D}^t$ and $\mathcal{T}^s = \mathcal{T}^t$, traditional ML methods can be used to solve the problem, where \mathcal{D}^s becomes the training set and \mathcal{D}^t the test set.

When this assumption does not hold, *i.e.* $\mathcal{D}^t \neq \mathcal{D}^s$ or $\mathcal{T}^t \neq \mathcal{T}^s$, the models trained on \mathcal{D}^s might perform poorly on \mathcal{D}^t , or they are not applicable directly if $\mathcal{T}^t \neq \mathcal{T}^s$. When the source domain is somewhat related to the target, it is possible to exploit the related information from $\{\mathcal{D}^s, \mathcal{T}^s\}$ to learn $P(\mathbf{Y}^t|\mathbf{X}^t)$. This process is known as *transfer learning* (TL).

We distinguish between *homogeneous TL*, where the source and target are represented in the same the feature space, $\mathcal{X}^t = \mathcal{X}^s$, with $P(\mathbf{X}^t) \neq P(\mathbf{X}^s)$ due to domain shift, and *heterogeneous TL* where the source and target data can have different representations, $\mathcal{X}^t \neq \mathcal{X}^s$ (or they can even be of different modalities such as image *vs.* text).

Based on these definitions, [37] categorizes the TL approaches into three main groups depending on the different situations concerning source and target domains and the corresponding tasks. These are the inductive TL, transductive TL and unsupervised TL (see Figure 2). The *inductive TL* is the case where the target task is different but related to the source task, no matter whether the source and target domains are the same or not. It requires at least some labeled target instances to induce a predictive model for the target data. In the case of the *transductive TL*, the source and target tasks are the same, and either the source and target data representations are different ($\mathcal{X}^t \neq \mathcal{X}^s$) or the source and target distributions are different due to selection bias or distribution mismatch. Finally, the *unsupervised TL* refers to the case where both the domains and the tasks are different but somewhat related. In general, labels are not available neither for the source nor for the target, and the focus is on exploiting the (unlabeled) information in the source domain to solve unsupervised learning task in the target domain. These tasks include clustering, dimensionality reduction and density estimation [38, 39].

According to this classification, DA methods are transductive TL solutions, where it is assumed that the tasks are the same, *i.e.* $\mathcal{T}^t = \mathcal{T}^s$. In general they refer to a categorization task, where both the set of labels and the conditional distributions are assumed to be shared between the two domains, *i.e.* $\mathcal{Y}^s = \mathcal{Y}^t$ and $P(\mathbf{Y}|\mathbf{X}^t) = P(\mathbf{Y}|\mathbf{X}^s)$. However, the second assumption is rather strong and does not always hold in real-life applications. Therefore, the definition of domain adaptation is relaxed to the case where only the first assumption is required, *i.e.* $\mathcal{Y}^s = \mathcal{Y}^t = \mathcal{Y}$.

In the DA community, we further distinguish between the *unsupervised*² (US) case where the labels are available only for the source domain and the *semi-supervised* (SS) case where a small set of target examples are labeled.

² Note also that the unsupervised DA is not related to the unsupervised TL, for which no source labels are available and in general the task to be solved is unsupervised.

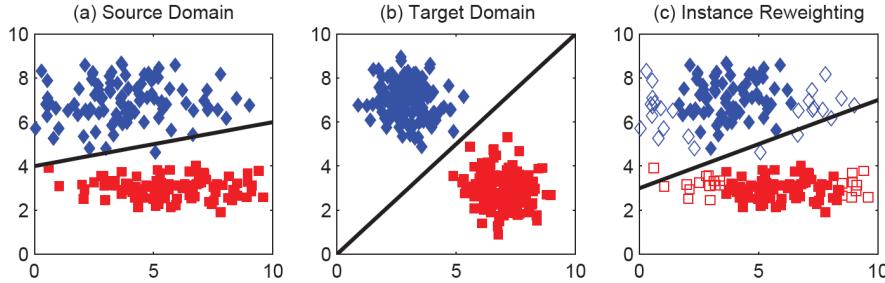


Fig. 3 Illustration of the effect of instance re-weighting samples on the source classifier. (Image: Courtesy to M. Long [40].)

3 Shallow domain adaptation methods

In this section, we review shallow DA methods that can be applied on vectorial visual features extracted from images. First, in Section 3.1 we survey homogeneous DA methods, where the feature representation for the source and target domains is the same, $\mathcal{X}^t = \mathcal{X}^s$ with $P(\mathbf{X}^t) \neq P(\mathbf{X}^s)$, and the tasks shared, $\mathcal{Y}^s = \mathcal{Y}^t$. Then, in Section 3.2 we discuss methods that can exploit efficiently several source domains. Finally in Section 3.3 we discuss the heterogeneous case, where the source and target data have different representations.

3.1 Homogeneous domain adaptation methods

Instance re-weighting methods. The DA case when we assume that the conditional distributions are shared between the two domains, *i.e.* $P(\mathbf{Y}|\mathbf{X}^s) = P(\mathbf{Y}|\mathbf{X}^t)$, is often referred to as *dataset bias* or *covariate shift* [41]. In this case, one could simply apply the model learned on the source to estimate $P(\mathbf{Y}|\mathbf{X}^t)$. However, as $P(\mathbf{X}^s) \neq P(\mathbf{X}^t)$, the source model might yield a poor performance when applied on the target set despite of the underlying $P(\mathbf{Y}|\mathbf{X}^s) = P(\mathbf{Y}|\mathbf{X}^t)$ assumption. The most popular early solutions proposed to overcome this to happen are based on instance re-weighting (see Figure 3 for an illustration).

To compute the weight of an instance, early methods proposed to estimate the ratio between the likelihoods of being a source or target example. This can be done either by estimating the likelihoods independently using a domain classifier [42] or by approximating directly the ratio between the densities with a Kullback-Leibler Importance Estimation Procedure [43, 44]. However, one of the most popular measure used to weight data instances, used for example in [45, 46, 14], is the Maximum Mean Discrepancy (MMD) [47] computed between the data distributions in the two domains.

The method proposed in [48] infers re-sampling weights through maximum entropy density estimation. [41] improves predictive inference under covariate shift by weighting the log-likelihood function. The Importance Weighted Twin Gaussian Processes [32] directly learns the importance weight function, without going through density estimation, by using the relative unconstrained least-squares importance fitting. The Selective Transfer Machine [31] jointly optimizes the weights as well as the classifier's parameters to preserve the discriminative power of the new decision boundary.

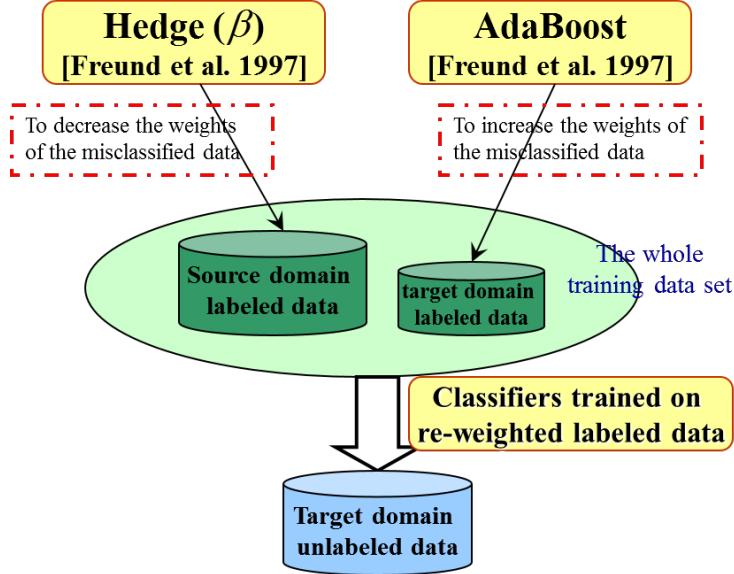


Fig. 4 Illustration of the TrAdaBoost method [49] where the idea is to decrease the importance of the misclassified source examples while focusing, as in AdaBoost [50], on the misclassified target examples. (Image: Courtesy to S. J. Pan).

The Transfer Adaptive Boosting (TrAdaBoost) [49], is an extension to AdaBoost³ [50], that iteratively re-weights both source and target examples during the learning of a target classifier. This is done by increasing the weights of miss-classified target instances as in the traditional AdaBoost, but decreasing the weights of miss-classified source samples in order to diminish their importance during the training process (see Figure 4). The TrAdaBoost was further extended by integrating dynamic updates in [51, 52].

Parameter adaptation methods. Another set of early DA methods, but which does not necessarily assume $P(\mathbf{Y}|\mathbf{X}^s) = P(\mathbf{Y}|\mathbf{X}^t)$, investigates different options to adapt the classifier trained on the source domain, e.g. an SVM, in order to perform better on the target domain⁴. Note that these methods in general require at least a small set of labeled target examples per class, hence they can only be applied in the semi-supervised DA scenario. As such, the Transductive SVM [53] that aims at decreasing the generalization error of the classification, by incorporating knowledge about the target data into the SVM optimization process. The Adaptive SVM (A-SVM) [54] progressively adjusts the decision boundaries of the source classifiers with the help of a set of so called perturbation functions built by exploiting predictions on the available labeled target examples (see Figure 5). The Domain Transfer SVM [55] simultaneously reduces the mismatch in the distributions (MMD) between two domains and learns a target decision function. The Adaptive Multiple Kernel Learning (A-MKL) [23] generalizes this by learning an adapted classifier based on multiple base

³ Code at <https://github.com/BoChen90/machine-learning-matlab/blob/master/TrAdaBoost.m>

⁴ The code for several methods, such as A-SVM, A-MKL, DT-MKL can be downloaded from <http://www.codeforge.com/article/248440>

Adaptive SVM [Yang et al. MM 2007]

$$f^s(\mathbf{x}) + \Delta f(\mathbf{x}) = f^t(\mathbf{x})$$

Source Classifier Perturbation function Target classifier

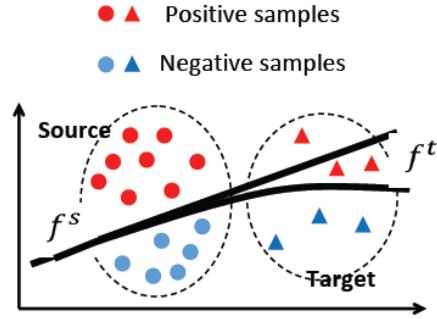


Fig. 5 Illustration of the Adaptive SVM [54], where a set of so called perturbation functions Δ_f are added to the source classifier f^s to progressively adjusts the decision boundaries of f^s for the target domain. (Courtesy to D. Xu).

kernels and the pre-trained average classifier. The model minimizes jointly the structural risk functional and the mismatch between the data distributions (MMD) of the two domains.

The domain adaptation SVM (DASVM) [56] exploits within the semi-supervised DA scenario both the transductive SVM [53] and its extension, the progressive transductive SVM [57]. The cross-domain SVM, proposed in [58], constrains the impact of source data to the k-nearest neighbors (similarly to the spirit of the Localized SVM [59]). This is done by down-weighting support vectors from the source data that are far from the target samples.

Feature augmentation. One of the simplest method for DA was proposed in [60], where the original representation \mathbf{x} is augmented with itself and a vector of the same size filled with zeros as follows: the source features become $\begin{bmatrix} \mathbf{x}^s \\ \mathbf{x}^s \end{bmatrix}$ and target features $\begin{bmatrix} \mathbf{x}^t \\ 0 \end{bmatrix}$. Then an SVM is trained on these augmented features to figure out which parts of the representation is shared between the domains and which are the domain specific ones.

The idea of feature augmentation is also behind the Geodesic Flow Sampling (GFS) [61, 62] and the Geodesic Flow Kernel (GFK) [18, 63], where the domains are embedded in d -dimensional linear subspaces that can be seen as points on the Grassman manifold corresponding to the collection of all d -dimensional subspaces. In the case of GFS [61, 62], following the geodesic path between the source and target domains, representations, corresponding to intermediate domains, are sampled gradually and concatenated (see illustration in Figure 6). Instead of sampling, GFK⁵ [18, 63], extends GFS to the infinite case, proposing a kernel that makes the solution equivalent to integrating over all common subspaces lying on the geodesic path.

A more generic framework, proposed in [62], accommodates domain representations in high-dimensional Reproducing Kernel Hilbert Space (RKHS) using kernel methods and low-dimensional manifold representations corresponding to Laplacian Eigenmaps. The approach described in [64] was inspired by the manifold-based incremental learning framework in [61]. It generates a set of intermediate dictionaries which smoothly connect the source and target domains. This is done by decomposing the target data with the current intermediate domain dictionary updated with a reconstruction residue estimated on the target. Concatenating these

⁵ Code available at http://www-scf.usc.edu/~boqinggo/domain_adaptation/GFK_v1.zip

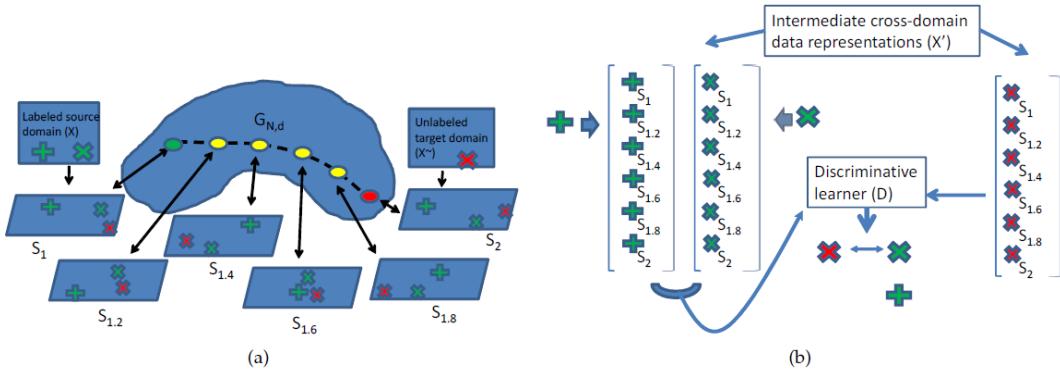


Fig. 6 The GFS samples between source S_1 and target S_2 on the geodesic path intermediate domains $S_{1,i}$ that can be seen as cross-domain data representations. (Courtesy to R. Gopalan [61].)

intermediate representations enables learning a better cross domain classifier.

These methods exploit *intermediate cross-domain* representations that are built without the use of class labels. Hence, they can be applied in both, the US and SS, scenarios. These cross-domain representations are then used either to train a discriminative classifier [62] using the available labeled set (only from the source or from both domains), or to label the target instances using nearest neighbor search in the kernel space [18, 63].

Feature space alignment. Instead of augmenting the features, other methods tries to align the source features with the target ones. As such, the Subspace Alignment (SA) [19] learns an alignment between the source subspace obtained by PCA and the target PCA subspace, where the PCA dimensions are selected by minimizing the Bregman divergence between the subspaces. Its advantage is its simplicity, as shown in Algorithm 1. Similarly, the linear Correlation Alignment (CORAL) [21] can be written in few lines of MATLAB code as illustrated in Algorithm 2. The method minimizes the domain shift by using the second-order statistics of the source and target distributions. The main idea is a whitening of the source data using its covariance followed by a "re-coloring" using the target covariance matrix.

As an alternative to feature alignment, a large set of feature transformation methods were proposed with the objective to find a projection of the data into a latent space such that the discrepancy between the source and target distributions is decreased. Note that the projections can be shared between the domains or they can be domain specific projections. In the latter case we talk about asymmetric feature transformation. Furthermore, when the transformation learning procedure uses no class labels, the method is called *unsupervised* feature transformation and when the transformation is learned by exploiting class labels (only from the source or also from the target when available) it is referred to as *supervised* feature transformation.

Unsupervised feature transformation. One of the first such DA method is the Transfer Component Analysis (TCA) [14] that proposes to discover common latent features having the same marginal distribution across the source and target domains, while maintaining the intrinsic structure (local geometry of the data manifold) of the original domain by a smoothness term.

Algorithm 1: Subspace Alignment (SA) [19]**Input:** Source data \mathbf{X}^s , target data \mathbf{X}^t , subspace dimension d 1: $\mathbf{P}_s \leftarrow PCA(\mathbf{X}^s, d)$, $\mathbf{P}_t \leftarrow PCA(\mathbf{X}^t, d)$;2: $\mathbf{X}_a^s = \mathbf{X}^s \mathbf{P}_s \mathbf{P}_s^\top \mathbf{P}_t$, $\mathbf{X}_a^t = \mathbf{X}^t \mathbf{P}_t$;**Output:** Aligned source, \mathbf{X}_a^s and target, \mathbf{X}_a^t data.**Algorithm 2:** Correlation Alignment (CORAL) [21]**Input:** Source data \mathbf{X}^s , target data \mathbf{X}^t 1: $\mathbf{C}_s = cov(\mathbf{X}^s) + eye(size(\mathbf{X}^s, 2))$, $\mathbf{C}_t = cov(\mathbf{X}^t) + eye(size(\mathbf{X}^t, 2))$ 2: $\mathbf{X}_w^s = \mathbf{X}^s * \mathbf{C}_s^{-1/2}$ (whitening), $\mathbf{X}_a^s = \mathbf{X}_w^s * \mathbf{C}_t^{-1/2}$ (re-coloring)**Output:** Source data \mathbf{X}_a^s adjusted to the target.

Instead of restricting the discrepancy to a simple distance between the sample means in the lower-dimensional space, Baktashmotagh *et al.* [65] propose the Domain Invariant Projection⁶ (DIP) approach that compares directly the distributions in the RKHS while constraining the transformation to be orthogonal. They go a step further in [66] and based on the fact that probability distributions lie on a Riemannian manifold, propose the Statistically Invariant Embedding⁷ (SIE) that uses the Hellinger distance on this manifold to compare kernel density estimates between of the source and target data. Both the DIP and SIE, involve non-linear optimizations and are solved with the conjugate gradient algorithm [67].

The Transfer Sparse Coding⁸ (TSC) [68] learns robust sparse representations for classifying cross-domain data accurately. To bring the domains closer, the distances between the sample means for each dimensions of the source and the target is incorporated into the objective function to be minimized. The Transfer Joint Matching⁹ (TJM) [40] learns a non-linear transformation between the two domains by minimizing the distance between the empirical expectations of source and target data distributions integrated within a kernel embedding. In addition, to put less emphasis on the source instances that are irrelevant to classify the target data, instance re-weighting is employed.

The feature transformation proposed by in [12] exploits the correlation between the source and target set to learn a robust representation by reconstructing the original features from their noised counterparts. The method, called Marginalized Denoising Autoencoder (MDA), is based on a quadratic loss and a drop-out noise level that factorizes over all feature dimensions. This allows the method to avoid explicit data corruption by marginalizing out the noise and to have a closed-form solution for the feature transformation. Note that it is straightforward to stack together several layers with optional non-linearities between layers to obtain a multi-layer network with the parameters for each layer obtained in a single forward pass (see Algorithm 3).

In general, the above mentioned methods learn the transformation without using any class label. After projecting the data in the new space, any classifier trained on the source set can be used to predict labels for the target data. The model often works even better if in addition a small set of the target examples are hand-labeled (SS adaptation). The class labels can also be used to learn a better transformation. Such methods,

⁶ Code at https://drive.google.com/uc?export=download&id=0B9_PW9TCpxT0c292bW1RaWtXRHc⁷ Code at https://drive.google.com/uc?export=download&id=0B9_PW9TCpxT0SEdMQ1pCNzdZekU⁸ Code at <http://ise.thss.tsinghua.edu.cn/~mlong/doc/transfer-sparse-coding-cvpr13.zip>⁹ Code at <http://ise.thss.tsinghua.edu.cn/~mlong/doc/transfer-joint-matching-cvpr14.zip>

Algorithm 3: Stacked Marginalized Denoising Autoencoder (sMDA) [12].

Input: Source data \mathbf{X}^s , target data \mathbf{X}^t
Input: Parameters: p (noise level), ω (regularizer) and k (number of stacked layers)
1: $\mathbf{X} = [\mathbf{X}^s, \mathbf{X}^t]$, $\mathbf{S} = \mathbf{X}^\top \mathbf{X}$, and $\mathbf{X}_0 = \mathbf{X}$;
2: $\mathbf{P} = (1 - p)\mathbf{S}$ and $\mathbf{Q} = (1 - p)^2\mathbf{S} + p(1 - p)\text{diag}(\mathbf{S})$
3: $\mathbf{W} = (\mathbf{Q} + \omega\mathbf{I}_D)^{-1}\mathbf{P}$.
4: (Optionally), stack K layers with $\mathbf{X}_{(k)} = \tanh(\mathbf{X}_{(k-1)}\mathbf{W}^{(k)})$.
Output: Denoised features \mathbf{X}_k .

called supervised feature transformation based DA methods, to learn the transformation exploit class labels, either only from the source or also from the target (when available). When only the source class labels are exploited, the method can still be applied to the US scenario, while methods using also target labels are designed for the SS case.

Supervised feature transformation. Several unsupervised feature transformation methods, cited above, have been extended to capitalize on class labels to learn a better transformation. Among these extensions, we can mention the Semi-Supervised TCA [14, 69] where the objective function that is minimized contains a label dependency term in addition to the distance between the domains and the manifold regularization term. The label dependency term has the role of maximizing the alignment of the projections with the source labels and, when available, target labels.

Similarly, in [70] a quadratic regularization term, relying on the pretrained source classifier, is added into the MDA framework [12], in order to keep the denoised source data well classified. Moreover, the domain denoising and cross-domain classifier can be learned jointly by iteratively solving a Sylvester linear system to estimate the transformation and a linear system to get the classifier in closed form¹⁰.

To take advantage of class labels, the distance between each source sample and its corresponding class means is added as regularizer into the DIP [65] respectively SIE model [66]. This term encourages the source samples from the same class to be clustered in the latent space. The Adaptation Regularization based Transfer Learning¹¹ [71] performs DA by optimizing simultaneously the structural risk functional, the joint distribution matching between domains and the manifold consistency. The Max-Margin Domain Transform¹² [72] optimizes both the transformation and classifier parameters jointly, by introducing an efficient cost function based on the misclassification loss.

Another set of methods extend marginal distribution discrepancy minimization to conditional distribution involving data labels from the source and class predictions from the target. Thus, [73] proposes an adaptive kernel approach that maps the marginal distribution of the target and source sets into a common kernel space, and use a sample selection strategy to draw conditional probabilities between the two domains closer. The Joint Distribution Adaptation¹³ [20] jointly adapts the marginal distribution through a principled (PCA based) dimensionality reduction procedure and the conditional distribution between the domains.

¹⁰ Code at https://github.com/sclincha/xrce_msda_da_regularization

¹¹ Code at <http://ise.thss.tsinghua.edu.cn/~mlong/doc/adaptation-regularization-tkde14.zip>

¹² Code at https://cs.stanford.edu/~jhoffman/code/Hoffman_ICLR13_MMDT_v3.zip

¹³ Code at <http://ise.thss.tsinghua.edu.cn/~mlong/doc/joint-distribution-adaptation-iccv13.zip>

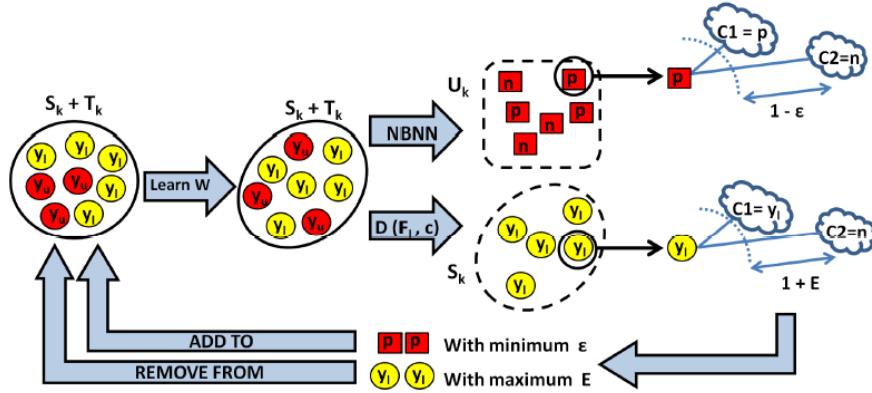


Fig. 7 The NBNN-DA adjusts the image-to-class distances by tuning the per class metrics and iteratively making the metric progressively more suitable for the target. (Image: Courtesy to T. Tommasi [61])

Metric learning based feature transformation. These methods are particular supervised feature transformation methods that involve that at least a limited set of target labels are available, and they use metric learning techniques to bridge the relatedness between the source and target domains. Thus, [74] proposes distance metric learning with either log-determinant or manifold regularization to adapt face recognition models between subjects. [17] uses the Information-Theoretic Metric Learning from [75] to define a common distance metric across different domains. This method was further extended in [76] by incorporating non-linear kernels, which enable the model to be applicable to the heterogeneous case (*i.e.* different source and target representations).

The metric learning for Domain Specific Class Means (DSCM) [77] learns a transformation of the feature space which, for each instance minimizes the weighted soft-max distances to the corresponding domain specific class means. This allows in the projected space to decrease the intraclass and to increase the interclass distances (see also Figure 10). This was extended with an active learning component by the Self-adaptive Metric Learning Domain Adaptation (SaML-DA) [77] framework, where the target training set is iteratively increased with labels predicted with DSCM and used to refine the current metric. SaML-DA was inspired by the Naive Bayes Nearest Neighbor based Domain Adaptation¹⁴ (NBNN-DA) [78] framework, which combines metric learning and NBNN classifier to adjust the instance-to-class distances by progressively making the metric more suitable for the target domain (see Figure 7). The main idea behind both methods, SaML-DA and NBNN-DA, is to replace at each iteration the most ambiguous source example of each class by the target example for which the classifier (DSCM respectively NNBA) is the most confident for the given class.

Local feature transformation. The previous methods learn a global transformation to be applied to each source and target example. In contrast, the Adaptive Transductive Transfer Machines (ATTM) [80] complements the global transformation with a sample-based transformation to refine the probability density function of the source instances assuming that the transformation from the source to the target domain is locally lin-

¹⁴ Code at <http://www.tatianatommasi.com/2013/DANBNNdemo.tar.gz>

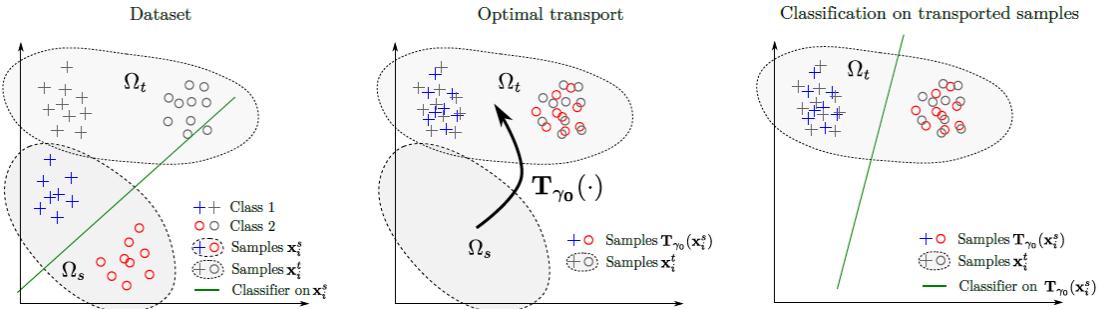


Fig. 8 The OTDA [79] consider a local transportation plan for each sample in the source domain to transport the training samples close to the target examples. (Image: Courtesy to N. Courty.)

ear. This is achieved by representing the target set by a Gaussian Mixture Model and learning an optimal translation parameter that maximizes the likelihood of the translated source as a posterior.

Similarly, the Optimal Transport for Domain Adaptation [79], considers a local transportation plan for each source example. The model can be seen as a graph matching problem, where the final coordinates of each sample are found by mapping the source samples to the target ones, whilst respecting the marginal distribution of the target domain (see Figure 8). To exploit class labels, a regularization term with group-lasso is added inducing, on one hand, group sparsity and, on another hand, constraining source samples of the same class to remain close during the transport.

Landmark selection. In order to improve the feature learning process, several methods have been proposed with the aim of selecting the most relevant instances from the source, so-called landmark examples, to be used to train the adaptation model (see examples in Figure 9). Thus, [63] proposes to minimize a variant of the MMD to identify good landmarks by creating a set of auxiliary tasks that offer multiple views of the original problem¹⁵. The Statistically Invariant Sample Selection [66], uses the Hellinger distance on the statistical manifold instead of MMD. The selection is forced to keep the proportions of the source samples per class the same as in the original data. Contrariwise to these approaches, the Multi-scale Landmark Selection¹⁶ [81] does not require any class labels. It takes each instance independently and considers it as being a good candidate if the Gaussian distributions of the source examples and of the target points centered on the instance are similar over a set of different scales (Gaussian variances).

Note that the landmark selection process, although strongly related to instance re-weighting methods with binary weights, can be rather seen as data preprocessing and hence complementary to the adaptation process.

¹⁵ Code at http://www-scf.usc.edu/~boqinggo/domain_adaptation/landmark_v1.zip

¹⁶ Code at <http://home.heeere.com/data/cvpr-2015/LSSA.zip>

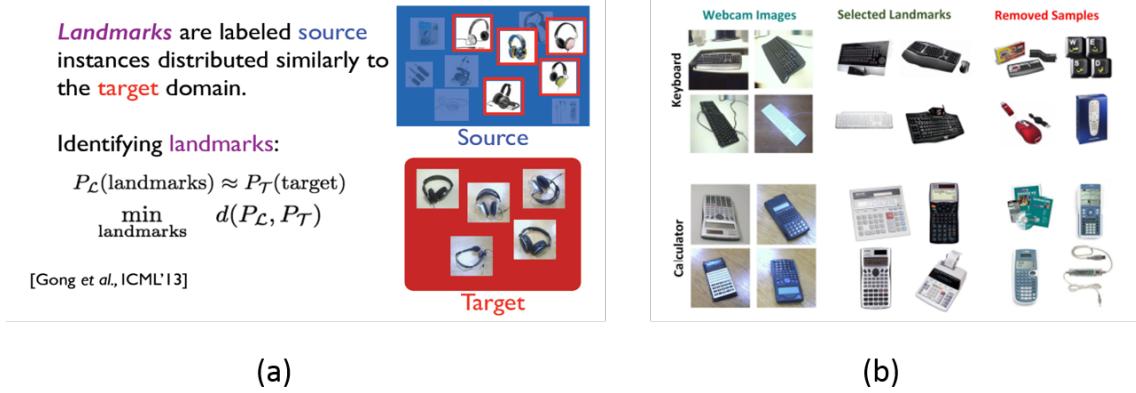


Fig. 9 Landmarks selected for the task *amazon versus webcam* using the popular Office31 dataset [17] with (a) MMD [63] and (b) the Hellinger distance on the statistical manifold [66].

3.2 Multi-source domain adaptation

Most of the above mentioned methods were designed for a single source *vs.* target case. When multiple sources are available, they can be concatenated to form a *single* source set, but because the possible shift between the different source domains, this might not be always a good option. Alternatively, the models built for each *source-target* pair (or their results) can be combined to make a final decision. However, a better option might be to build multi-source DA models which, relying only on the *a priori known* domain labels, are able to exploit the specificity of each source domain.

Such methods are the Feature Augmentation (FA) [60] and the A-SVM [54], already mentioned in Section 3.1, both exploiting naturally the multi-source aspect of the dataset. Indeed in the case of FA, extra feature sets, one for each source domain, concatenated to the representations, allow to learn source specific properties shared between a given source and the target. The A-SVM uses an ensemble of source specific auxiliary classifiers to adjust the parameters of the target classifier.

Similarly, the Domain Adaptation Machine [82] leverages a set of source classifiers by the integration of domain-dependent regularizer term which is based on a smoothness assumption. The model forces the target classifier to share similar decision values with the relevant source classifiers on the unlabeled target instances. The Conditional Probability based Multi-source Domain Adaptation (CP-MDA) approach [83] extends the above idea by adding weight values for each source classifier based on conditional distributions. The DSCM proposed in [77] relies on domain specific class means both to learn the metric but also to predict the target class labels (see illustration in Figure 10). The domain regularization and classifier based regularization terms of the extended MDA [70] are both sums of source specific components.

The Robust DA via Low-Rank Reconstruction (RDALRR) [84] transforms each source domain into an intermediate representation such that the transformed samples can be linearly reconstructed from the target ones. Within each source domain, the intrinsic relatedness of the reconstructed samples is imposed by using a low-rank structure where the outliers are identified using sparsity constraints. By enforcing different source domains to have jointly low ranks, a compact source sample set is formed with a distribution close to the target domain (see Figure 11).

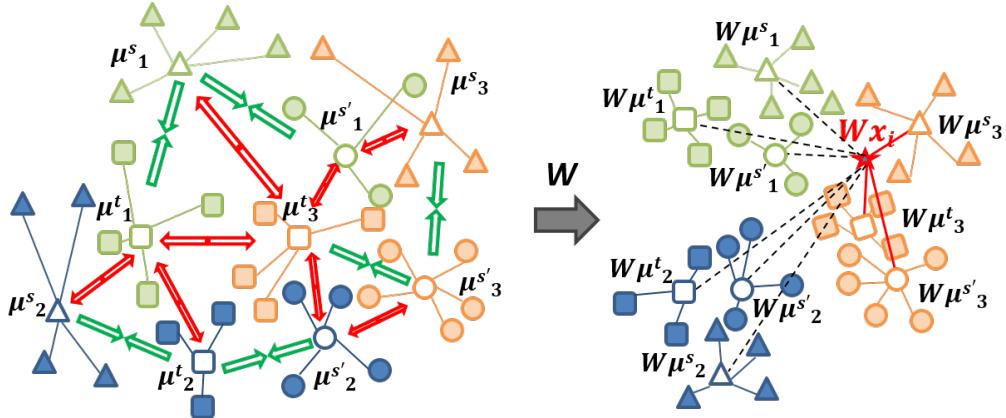


Fig. 10 Metric learning for the DSCM classifier, where $\mu_{c_i}^s$ and $\mu_{c_i}^{s'}$ represent source specific class means and $\mu_{c_i}^t$ class means in the target domain. The feature transformation \mathbf{W} is learned by minimizing for each sample the weighted soft-max distances to the corresponding domain specific class means in the projected space.

To better take advantage of having multiple source domains, extensions to methods previously designed for a single source *vs.* target case were proposed in [62, 85, 86, 87]. Thus, [62] describes a multi-source version of the GFS [61], which was further extended in [85] to the Subspaces by Sampling Spline Flow approach. The latter uses smooth polynomial functions determined by splines on the manifold to interpolate between different source and the target domain. [86] combines¹⁷ constrained clustering algorithm, used to identify automatically source domains in a large data set, with a multi-source extension of the Asymmetric Kernel Transform [76]. [87] efficiently extends the TrAdaBoost [49] to multiple source domains.

Source domain weighting. When multiple sources are available, it is desired to select those domains that provide the best information transfer and to remove the ones that have more likely negatively impact on the final model. Thus, to down-weight the effect of less related source domains, in [88] first the available labels are propagated within clusters obtained by spectral clustering and then to each source cluster a Supervised Local Weight (SLW) is assigned based on the percentage of label matches between predictions made by a source model and those made by label propagation.

In the Locally Weighted Ensemble framework [88], the model weights are computed as a similarity between the local neighborhood graphs centered on source and target instances. The CP-MDA [83], mentioned above, uses a weighted combination of source learners, where the weights are estimated as a function of conditional probability differences between the source and target domains. The Rank of Domain value defined in [18] measures the relatedness between each source and target domain as the KL divergences between data distributions once the data is projected into the latent subspace. The Multi-Model Knowledge Transfer [89] minimizes the negative transfer by giving higher weights to the most related linear SVM source classifiers. These weights are determined through a leave one out learning process.

¹⁷ Code at https://cs.stanford.edu/~jhoffman/code/hoffman_latent_domains_release_v2.zip

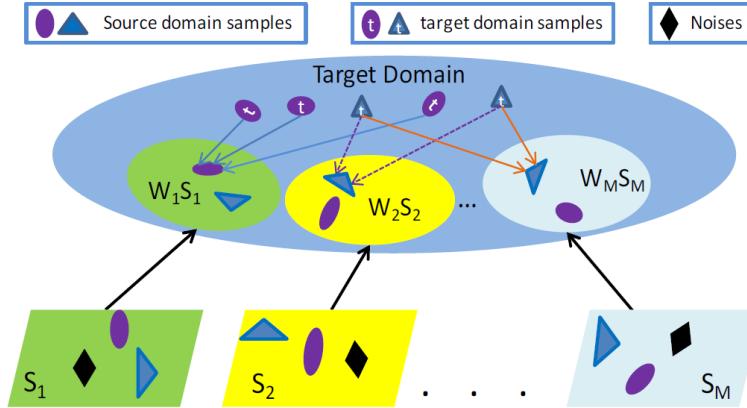


Fig. 11 The RDALRR [84] transforms each source domain into an intermediate representation such that the transformed samples can be linearly reconstructed from the target samples. (Image: Courtesy to I.H. Jhuo.)

3.3 Heterogeneous domain adaptation

Heterogeneous transfer learning (HTL) refers to the setting where the representation spaces are different for the source and target domains ($\mathcal{X}^t \neq \mathcal{X}^s$ as defined in Section 2). As a particular case, when the tasks are assumed to be the same, *i.e.* $\mathcal{Y}^s = \mathcal{Y}^t$, we refer to it as *heterogeneous domain adaptation* (HDA).

Both HDA and HTL are strongly related to multi-view learning [90, 91], where the presence of multiple information sources gives an opportunity to learn better representations (features) by analyzing the views simultaneously. This makes possible to solve the task when not all the views are available. Such situations appear when processing simultaneously audio and video [92], documents containing both image and text (*e.g.* web pages or photos with tags or comments) [93, 94, 95], images acquired with depth information [96], *etc.* We can also have multi-view settings when the views have the same modalities (textual, visual, audio), such as in the case of parallel text corpora in different languages [97, 98], photos of the same person taken across different poses, illuminations and expressions [27, 29, 99, 100].

Multi-view learning assumes that at training time for the same data instance multiple views from complementary information sources are available (*e.g.* a person is identified by photograph, fingerprint, signature or iris). Instead, in the case of HTL and HDA, the challenge comes from the fact that we have one view at training and another one at test time. Therefore, one set of methods proposed to solve HDA relies on some multi-view auxiliary data¹⁸ to bridge the gap between the domains (see Figure 12).

Methods relying on auxiliary domains. These methods principally exploit feature co-occurrences (*e.g.* between words and visual features) in the multi-view auxiliary domain. As such, the Transitive Transfer Learning [101] selects an appropriate domain from a large data set guided by domain complexity and, the distribution differences between the original domains (source and target) and the selected one (auxiliary).

¹⁸ When the bridge is to be done between visual and textual representations, a common practice is to crawl the Web for pages containing both text and images in order to build such intermediate multi-view data.

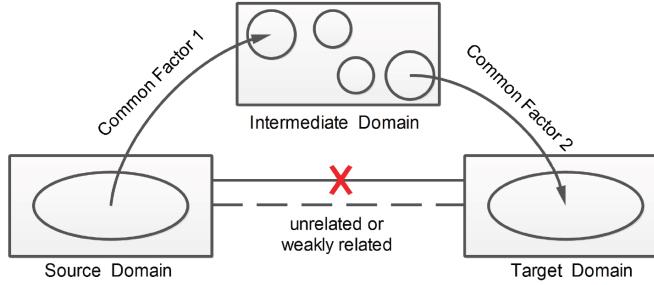


Fig. 12 Heterogeneous DA through an intermediate domain allowing to bridge the gap between features representing the two domains. For example, when the source domain contains text and the target images, the intermediate domain can be built from a set of crawled Web pages containing both text and images. (Image courtesy B. Tan [101]).

Then, using Non-negative Matrix Tri-factorization [102], feature clustering and label propagation is performed simultaneously through the intermediate domain.

The Mixed-Transfer approach [103] builds a joint transition probability graph of mixed instances and features, considering the data in the source, target and intermediate domains. The label propagation on the graph is done by a random walk process to overcome the data sparsity. In [104] the representations of the target images are enriched with semantic concepts extracted from the intermediate data¹⁹ through a Collective Matrix Factorization [105].

[106] proposes to build a translator function²⁰ between the source and target domain by learning directly the product of the two transformation matrices that map each domain into a common (hypothetical) latent topic built on the co-occurrence data. Following the principle of parsimony, they encode as few topics as possible in order to be able to match text and images. The semantic labels are propagated from the labeled text corpus to unlabeled new images by a cross-domain label propagation mechanism using the built translator. In [107] the co-occurrence data is represented by the principal components computed in each feature space and a Markov Chain Monte Carlo [108] is employed to construct a directed cyclic network where each node is a domain and each edge weight represents the conditional dependence between the corresponding domains defined by the transfer weights.

[109] studies online HDA, where offline labeled data from a source domain is transferred to enhance the online classification performance for the target domain. The main idea is to build an offline classifier based on heterogeneous similarity using labeled data from a source domain and unlabeled co-occurrence data collected from Web pages and social networks (see Figure 13). The online target classifier is combined with the offline source classifier using Hedge weighting strategy, used in Adaboost [50], to update their weights for ensemble prediction.

Instead of relying on external data to bridge the data representation gap, several HDA methods exploit directly the data distribution in the source and target domains willing to remove simultaneously the gap between the feature representations and minimizing the data distribution shift. This is done by learning either a

¹⁹ Code available at <http://www.cse.ust.hk/%7Eyinz/htl4ic.zip>

²⁰ Code available at http://www.ifp.illinois.edu/%7Eqi4/TTI_release_v1.zip

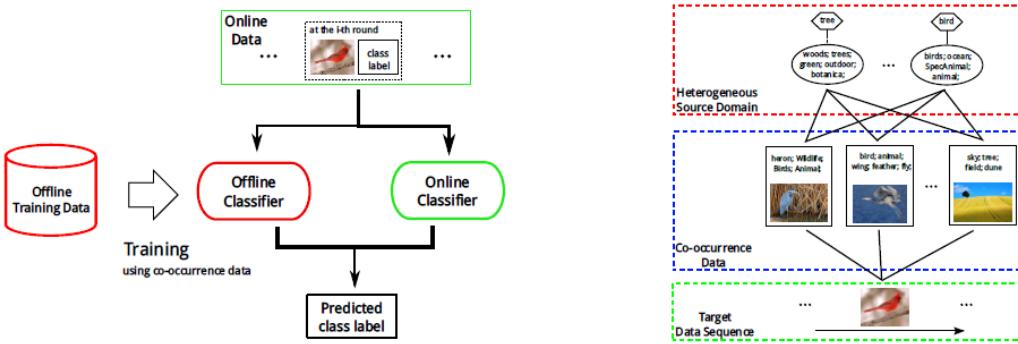


Fig. 13 Combining the online classifier with the offline classifier (right) and transfer the knowledge through co-occurrences data in the heterogeneous intermediate domain (left). (Image: Courtesy to Y. Yan [109])

projection for each domain into a domain-invariant common latent space, referred to as *symmetric transformation* based HDA²¹, or a transformation from the source space towards the target space, called *asymmetric transformation* based HDA. These approaches require at least a limited amount of labeled target examples (semi-supervised DA).

Symmetric feature transformation. The aim of symmetric transformation based HDA approaches is to learn projections for both the source and target spaces into a common latent (embedding) feature space better suited to learn the task for the target. These methods are related, on one hand, to the feature transformation based homogeneous DA methods described in Section 3.1 and, on another hand, to multi-view embedding [93, 110, 99, 111, 112, 113], where different views are embedded in a common latent space. Therefore, several DA methods originally designed for the homogeneous case, have been inspired by the multi-view embedding approaches and extended to heterogeneous data.

As such, the Heterogeneous Feature Augmentation²² (HFA) [114], prior to data augmentation, embeds the source and target into a common latent space (see Figure 15). In order to avoid the explicit projections, the transformation metrics are computed by the minimization of the structural risk functional of SVM expressed as a function of these projection matrices. The final target prediction function is computed by an alternating optimization algorithm to simultaneously solve the dual SVM and to find the optimal transformations. This model was further extended in [115], where each projection matrix is decomposed into a linear combination of a set of rank-one positive semi-definite matrices and they are combined within a Multiple Kernel Learning approach.

The Heterogeneous Spectral Mapping [116] unifies different feature spaces using spectral embedding where the similarity between the domains in the latent space is maximized with the constraint to preserve the original structure of the data. Combined with a source sample selection strategy, a Bayesian-based approach is applied to model the relationship between the different output spaces.

²¹ These methods can be used even if the source and target data are represented in the same feature space, *i.e.* $\mathcal{X}^t = \mathcal{X}^s$. Therefore, it is not surprising that several methods are direct extensions of homogeneous DA methods described in Section 3.1.

²² Code available at https://sites.google.com/site/xyzliwen/publications/HFA_release_0315.rar

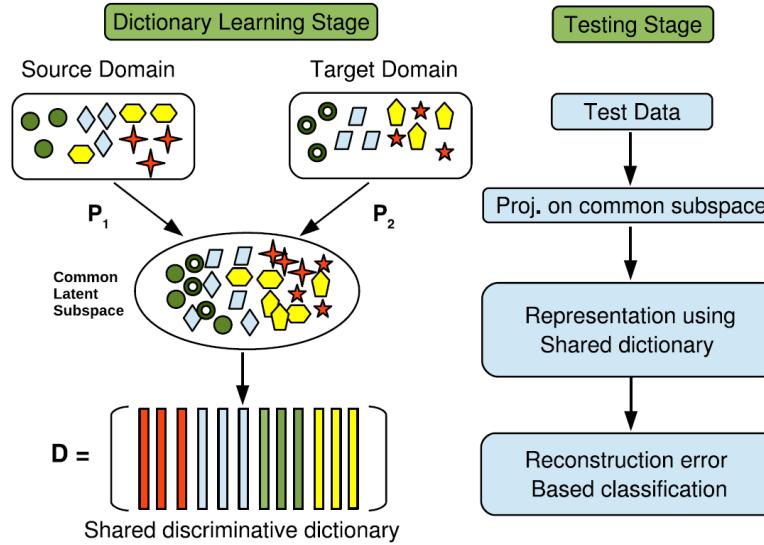


Fig. 14 The SDDL proposes to learn a dictionary in a latent common subspace while maintaining the manifold structure of the data. (Image: Courtesy to S. Shekhar [28])

[117] present a semi-supervised subspace co-projection method, which addresses heterogeneous multi-class DA. It is based on discriminative subspace learning and exploit unlabeled data to enforce an MMD criterion across domains in the projected subspace. They use Error Correcting Output Codes (ECOC) to address the multi-class aspect and to enhance the discriminative informativeness of the projected subspace. The Semi-supervised Domain Adaptation with Subspace Learning [118] jointly explores invariant low-dimensional structures across domains to correct data distribution mismatch and leverages available unlabeled target examples to exploit the underlying intrinsic information in the target domain.

To deal with both domain shift and heterogeneous data, the Shared Domain-adapted Dictionary Learning²³ (SDDL) [28] learns a class-wise discriminative dictionary in the latent projected space (see Figure 14). This is done by jointly learning the dictionary and the projections of the data from both domains onto a common low-dimensional space, while maintaining the manifold structure of data represented by sparse linear combinations of dictionary atoms.

The Domain Adaptation Manifold Alignment (DAMA) [119] models each domain as a manifold and creates a separate mapping function to transform the heterogeneous input space into a common latent space while preserving the underlying structure of each domain. This is done by representing each domains with a Laplacian that captures the closeness of the instances sharing the same label. The RDALRR [84], mentioned above (see also Figure 11), transforms each source domain into an intermediate representation such that the source samples linearly reconstructed from the target samples are enforced to be related to each other under a low-rank structure. Note that both DAMA and RDALRR are multi-source HDA approaches.

²³ Code available at <http://www.umiacs.umd.edu/~pvishalm/Codes/DomainAdaptDict.zip>

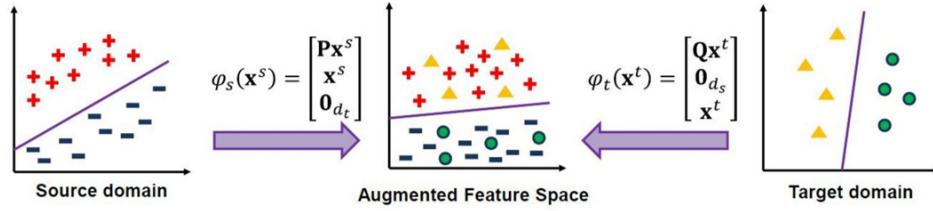


Fig. 15 The HFA [114] is seeking for an optimal common space while simultaneously learning a discriminative SVM classifier. (Image: Courtesy to Dong Xu.)

Asymmetric feature transformation. In contrast to symmetric transformation based HDA, these methods aim to learn a projection of the source features into the target space such that the distribution mismatch within each class is minimized. Such method is the Asymmetric Regularized Cross-domain Transformation²⁴ [76] that utilizes an objective function responsible for the domain invariant transformation learned in a non-linear Gaussian RBF kernel space. The Multiple Outlook MAPping algorithm [120] finds the transformation matrix by singular value decomposition process that encourage the marginal distributions within the classes to be aligned while maintaining the structure of the data. It requires a limited amount of labeled target data for each class to be paired with the corresponding source classes.

[10] proposes a sparse and class-invariant feature mapping that leverages the weight vectors of the binary classifiers learned in the source and target domains. This is done by considering the learning task as a Compressed Sensing [121] problem and using the ECOC scheme to generate a sufficient number of binary classifiers given the set of classes.

4 Deep domain adaptation methods

With the recent progress in image categorization due to deep convolutional architectures - trained in a fully supervised fashion on large scale annotated datasets, in particular on part of ImageNet [122] - allowed a significant improvement of the categorization accuracy over previous state-of-the art solutions. Furthermore, it was shown that features extracted from the activation layers of these deep convolutional networks can be re-purposed to novel tasks [123] even when the new tasks differ significantly from the task originally used to train the model.

Concerning domain adaptation, baseline methods without adaptation obtained using features generated by deep models²⁵ on the two most popular benchmark datasets Office (OFF31) [17] and Office+Caltech (OC10) [18] outperform by a large margin the shallow DA methods using the SURFBOV features originally provided with these datasets. Indeed, the results obtained with such Deep Convolutional Activation Features²⁶ (DeCAF) [123] even without any adaptation to the target are significantly better than the results

²⁴ Code available at http://vision.cs.uml.edu/code/DomainTransformsECCV10_v1.tar.gz

²⁵ Activation layers extracted from popular CNN models, such as AlexNet [124], VGGNET [125], ResNet [126] or GoogleNet [127].

²⁶ Code to extract features available at <https://github.com/UCBAIR/decaf-releas>

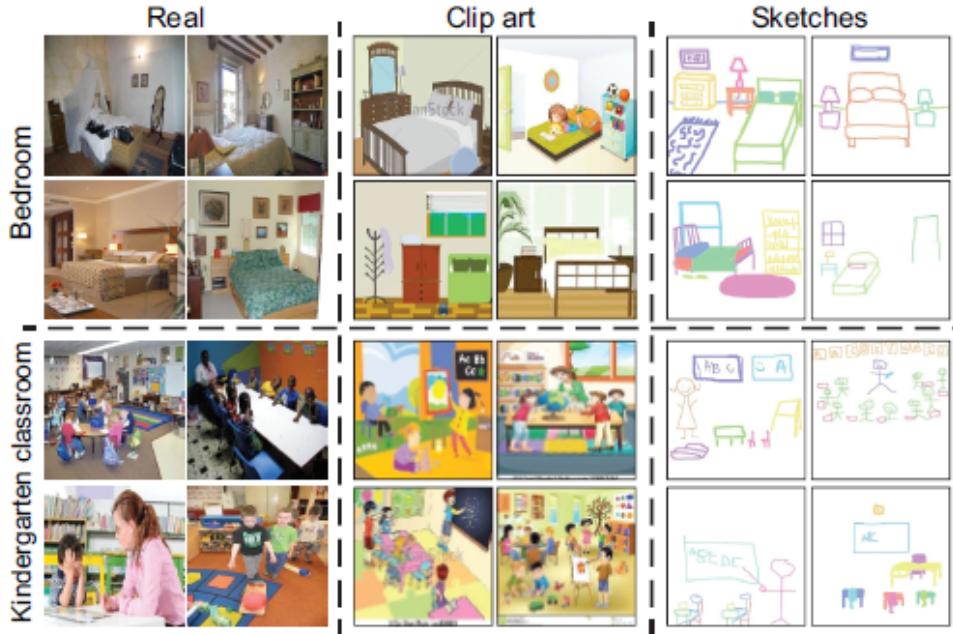


Fig. 16 Examples from the Cross-Modal Places Dataset (CMPlaces) dataset proposed in [3]. (Image: Courtesy to L. Castrejón.)

obtained with any DA method based on SURFBOV [128, 123, 21, 70]. As shown also in [129, 130], this suggests that deep neural networks learn more abstract and robust representations, encode category level information and remove, to a certain measure, the domain bias [123, 21, 70, 4].

Note however that in OFF31 and OC10 datasets the images remain relatively similar to the images used to train these models (usually datasets from the ImageNet Large-Scale Visual Recognition Challenge [122]). In contrast, if we consider category models between *e.g.* images and paintings, drawings, clip art or sketches (see see examples from the CMPlaces dataset²⁷ in Figure 16), the models have more difficulties to handle the domain differences [1, 131, 2, 3] and alternative solutions are necessary.

Solutions proposed in the literature to exploit deep models can be grouped into three main categories. The first group considers the CNN models to extract vectorial features to be used by the shallow DA methods. The second solution is to train or fine-tune the deep network on the source domain, adjust it to the new task, and use the model to predict class labels for target instances. Finally, the most promising methods are based on deep learning architectures designed for DA.

Shallow methods with deep features. The first, naive solution is to consider the deep network as feature extractor, where the activations of a layer or several layers of the deep architecture is considered as representation for the input image. These Deep Convolutional Activation Features (DeCAF) [123] extracted from both source and target examples can then be used within any shallow DA method described in Section 3. For example, Feature Augmentation [60], Max-Margin Domain Transforms [72] and Geodesic Flow Kernel [18]

²⁷ Dataset available at <http://projects.csail.mit.edu/cmplaces/>

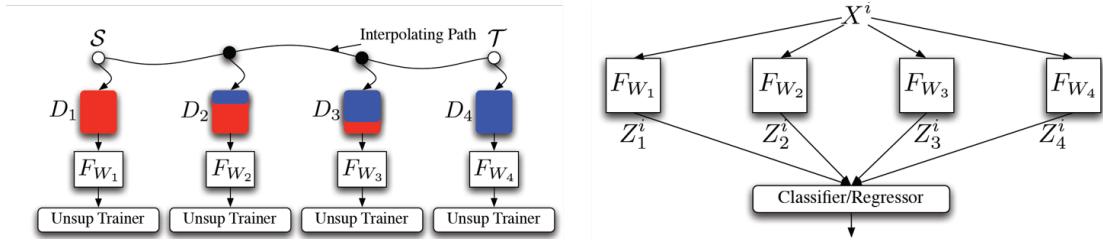


Fig. 17 The DLID model aims in interpolating between domains based on the amount of source and target data used to train each model. (Image courtesy S. Chopra [128]).

were applied to DECAF features in [123], Subspace Alignment [19] and Correlation Alignment in [21]. [70] experiments with DeCAF features within the extended MDA framework, while [4] explores various metric learning approaches to align deep features extracted from RGB face images (source) and NIR or sketches (target).

In general, these DA methods allow to further improve the classification accuracy compared to the baseline classifiers trained only on the source data with these DeCAF features [123, 21, 70, 4]. Note however that the gain is often relatively small and significantly lower than the gain obtained with the same methods when used with the SURFBOV features.

Fine-tuning deep CNN architectures. The second and most used solution is to fine-tune the deep network model on the new type of data and for the new task [132, 133, 134, 135]. But fine-tuning requires in general a relatively large amount of annotated data which is not available for the target domain, or it is very limited. Therefore, the model is in general fine-tuned on the source - augmented with, when available, the few labeled target instances - which allows in a first place to adjust the deep model to the new task²⁸, common between the source and target in the case of DA. This is fundamental if the targeted classes do not belong to the classes used to pretrain the deep model. However, if the domain difference between the source and target is important, fine-tuning the model on the source might over-fit the model for the source. In this case the performance of the fine-tuned model on the target data can be worse than just training the class prediction layer or as above, using the model as feature extractor and training a classifier²⁹ with the corresponding DeCAF features [128, 21].

4.1 DeepDA architectures

Finally, the most promising are the deep domain adaptation (DeepDA) methods that are based on deep learning architectures designed for domain adaptation. One of the first deep model used for DA is the Stacked Denoising Autoencoders [137] proposed to adapt sentiment classification between reviews of different products [13]. This model aims at finding common features between the source and target collections relying on denoising autoencoders. This is done by training a multi-layer neural network to reconstruct input data from partial random corruptions with backpropagation. The Stacked Marginalized Denoising Autoencoders [12]

²⁸ This is done by replacing the class prediction layer to correspond to the new set of classes.

²⁹ Note that the two approaches are equivalent when the layer preceding the class prediction layer are extracted.

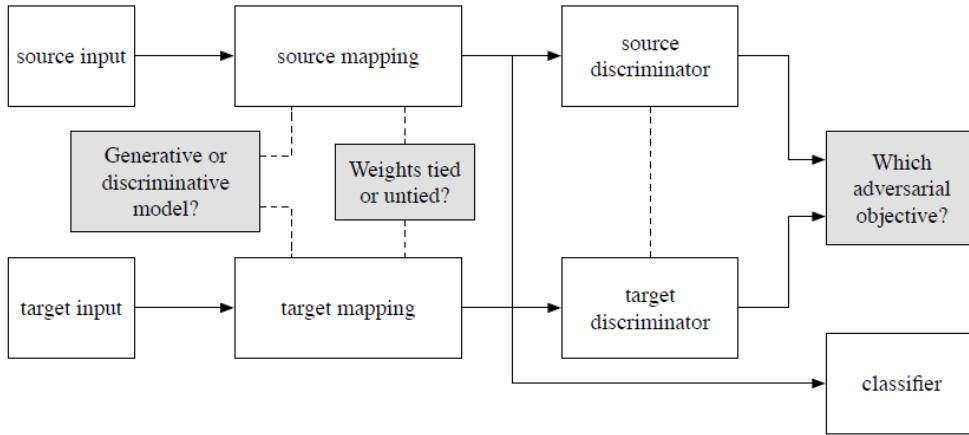


Fig. 18 Adversarial adaptation methods can be viewed as instantiations of the same framework with different choices regarding their properties [136] (Image courtesy E. Tzeng).

(see also in Section 3.1) is a variant of the SDA, where the random corruption is marginalized out and hence yields a unique optimal solution (feature transformation) computed in closed form between layers.

The Domain Adaptive Neural Network³⁰ [138] uses such denoising auto-encoder as a pretraining stage. To ensure that the model pretrained on the source continue to adapt to the target, the MMD is embedded as a regularization in the supervised backpropagation process (added to the cross-entropy based classification loss of the labels source examples).

The Deep Learning for Domain Adaptation [128], inspired by the intermediate representations on the geodesic path [18, 62], proposes a deep model based interpolation between domains. This is achieved by a deep nonlinear feature extractor trained in an unsupervised manner using the Predictive Sparse Decomposition [139] on intermediate datasets, where the amount of source data is gradually replaced by target samples.

[140] proposes a light-weight domain adaptation method, which, by using only a few target samples, analyzes and reconstructs the output of the filters that were found affected by the domain shift. The aim of the reconstruction is to make the filter responses given a target image resemble to the response map of a source image. This is done by simultaneously selecting and reconstructing the response maps of the bad filters using a Lasso based optimization with a KL-divergence measure that guides the filter selection process.

Most DeedDA methods follow a Siamese architectures [141] with two streams, representing the source and target models (see for example Figure 18), and are trained with a combination of a *classification loss* and a *discrepancy loss* [142, 143, 138, 144, 145] or an *adversarial loss*. The classification loss depends on the labeled source data. The discrepancy loss aims to diminish the shift between the two domains while the adversarial loss tries to encourage a common feature space through an adversarial objective with respect to

³⁰ Code available at <https://github.com/ghif/mtae>

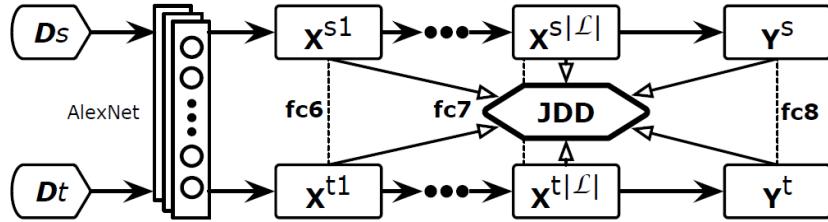


Fig. 19 The JAN [145] minimizes a joint distribution discrepancy of several intermediate layers including the soft prediction one. (Image courtesy M. Long).

a domain discriminator.

Discrepancy-based methods. These methods, inspired by the shallow feature space transformation approaches described in Section 3.1, uses in general a discrepancy based on MMD defined between corresponding activation layers of the two streams of the Siamese architecture. One of the first such method is the Deep Domain Confusion (DDC) [142] where the layer to be considered for the discrepancy and its dimension is automatically selected amongst a set of fine-tuned networks based on linear MMD between the source and the target. Instead of using a single layer and linear MMD, Long *et al.* proposed the Deep Adaptation Network³¹ (DAN) [143] that consider the sum of MMDs defined between several layers, including the soft prediction layer too. Furthermore, DAN explore multiple kernels for adapting these deep representations, which substantially enhances adaptation effectiveness compared to a single kernel method used in [138] and [142]. This was further improved by the Joint Adaptation Networks [145], which instead of the sum of marginal distributions (MMD) defined between different layers, consider the joint distribution discrepancies of these features.

The Deep CORAL [144] extends the shallow CORAL [21] method described in Section 3 to deep architectures³². The main idea is to learn a nonlinear transformation that aligns correlations of activation layers between the two streams. This idea is similarly to DDC and DAN except that instead of MMD the CORAL loss³³ (expressed by the distance between the covariances) is used to minimize discrepancy between the domains.

In contrast to the above methods, Rozantsev *et al.* [146] consider the MMD between the weights of the source respectively target models of different layers, where an extra regularizer term ensures that the weights in the two models remains linearly related.

Adversarial discriminative models. The aim of these models is to encourage domain confusion through an adversarial objective with respect to a domain discriminator. [136] proposes a unified view of existing adversarial DA methods by comparing them depending on the loss type, the weight sharing strategy between the two streams and, on whether they are discriminative or generative (see illustration in Figure 18). Amongst the discriminative models we have the model proposed in [148] using a confusion loss, the Ad-

³¹ Code available at <https://github.com/thumt/transfer-cafe>

³² Code available at <https://github.com/VisionLearningGroup/CORAL>

³³ Note that this loss can be seen as minimizing the MMD with a polynomial kernel.

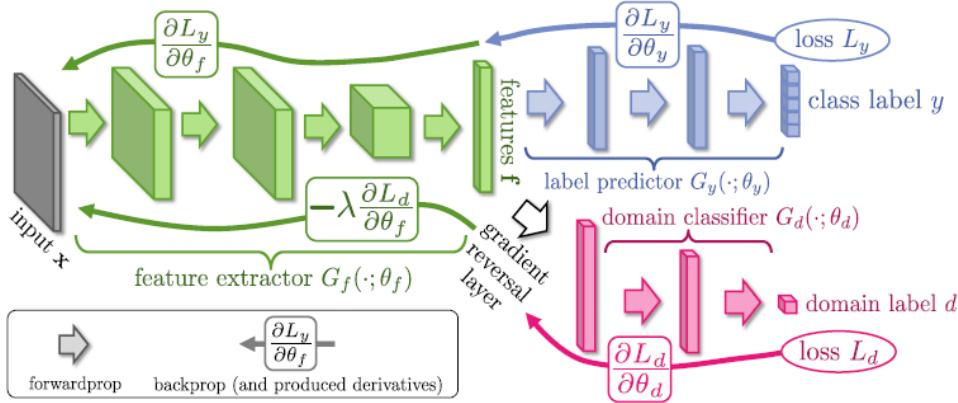


Fig. 20 The DANN architecture including a feature extractor (green) and a label predictor (blue), which together form a standard feed-forward architecture. Unsupervised DA is achieved by the gradient reversal layer that multiplies the gradient by a certain negative constant during the backpropagation-based training to ensure that the feature distributions over the two domains are made indistinguishable. (Image courtesy Y. Ganin [147]).

versarial Discriminative Domain Adaptation [136] that considers an inverted label GAN loss [149] and the Domain-Adversarial Neural Network [147] with a minimax loss. The generative methods, additionally to the discriminator, relies on a generator, which, in general, is a Generative Adversarial Network (GAN) [149].

The domain confusion based model³⁴ proposed in [148] considers a domain confusion objective, under which the mapping is trained with both unlabeled and sparsely labeled target data using a cross-entropy loss function against a uniform distribution. The model simultaneously optimizes the domain invariance to facilitate domain transfer and uses a soft label distribution matching loss to transfer information between tasks.

The Domain-Adversarial Neural Networks³⁵ (DANN) [147], integrates a gradient reversal layer into the standard architecture to promote the emergence of features that are discriminative for the main learning task on the source domain and indiscriminate with respect to the shift between the domains (see Figure 20). This layer is left unchanged during the forward propagation and its gradient reversed during backpropagation.

The Adversarial Discriminative Domain Adaptation [136] uses an inverted label GAN loss to split the optimization into two independent objectives, one for the generator and one for the discriminator. In contrast to the above methods, this model considers independent source and target mappings (unshared weights between the two streams) allowing domain specific feature extraction to be learned, where the target weights are initialized by the network pretrained on the source.

Adversarial generative models. These models combine the discriminative model with a generative component in general based on GANs [149]. As such, the Coupled Generative Adversarial Networks [150] consists of a tuple of GANs each corresponding to one of the domains. It learns a joint distribution of multi-domain images and enforces a weight sharing constraint to limit the network capacity.

³⁴ Code available at <https://github.com/erictzeng/caffe/tree/confusion>

³⁵ Code available at <https://github.com/ddtm/caffe/tree/grl>

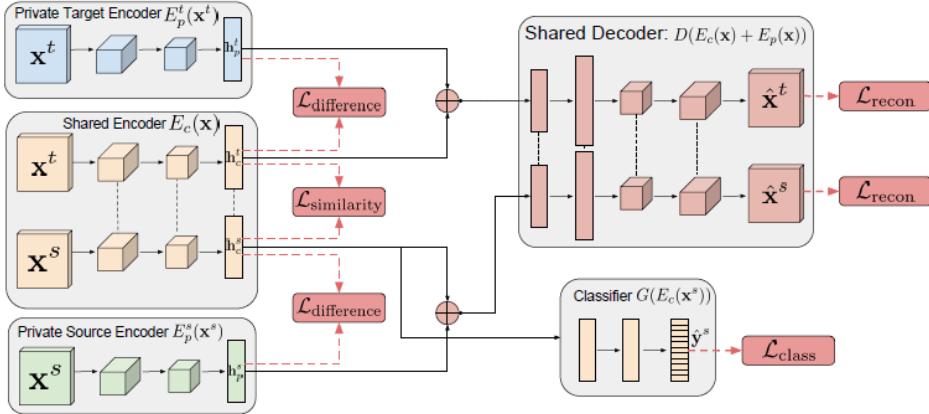


Fig. 21 The DSN architecture combines shared and domain specific encoders, which learns common and domain specific representation components respectively with a shared decoder that learns to reconstruct the input samples. (Image courtesy K. Bousmalis [153]).

The model proposed in [151] also exploit GANs with the aim to generate source-domain images such that they appear as if they were drawn from the target domain. Prior knowledge regarding the low-level image adaptation process, such as foreground-background segmentation mask, can be integrated in the model through content-similarity loss defined by a masked Pairwise Mean Squared Error [152] between the unmasked pixels of the source and generated images. As the model decouples the process of domain adaptation from the task-specific architecture, it is able to generalize also to object classes unseen during the training phase.

Data reconstruction (encoder-decoder) based methods. In contrast to the above methods, the Deep Reconstruction Classification Network³⁶ proposed in [154] combines the standard convolutional network for source label prediction with a deconvolutional network [155] for target data reconstruction. To jointly learn source label predictions and unsupervised target data reconstruction, the model alternates between unsupervised and supervised training. The parameters of the encoding are shared across both tasks, while the decoding parameters are separated. The data reconstruction can be viewed as an auxiliary task to support the adaptation of the label prediction.

The Domain Separation Networks (DSN) [153] introduces the notion of a private subspace for each domain, which captures domain specific properties, such as background and low level image statistics. A shared subspace, enforced through the use of autoencoders and explicit loss functions, captures common features between the domains. The model integrates a reconstruction loss using a shared decoder, which learns to reconstruct the input sample by using both the private (domain specific) and source representations (see Figure 21).

³⁶ Code available at <https://github.com/ghif/drcn>

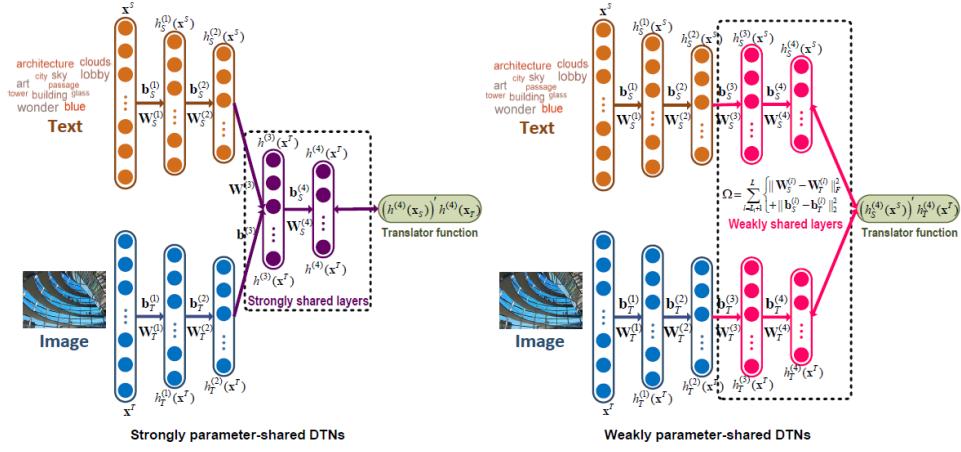


Fig. 22 The DTN architecture with strongly-shared and weakly-shared parameter layers. (Image courtesy X. Shu [157]).

Heterogeneous deepDA. Concerning heterogeneous or multi-modal deep domain adaptation, we can mention the Transfer Neural Trees [156] proposed to relate heterogeneous cross-domain data. It is a two stream network, one stream for each modality, where the weights in the latter stages of the network are shared. As the prediction layer, a Transfer Neural Decision Forest (Transfer-NDF) is used that performs jointly adaptation and classification.

The weakly-shared Deep Transfer Networks for Heterogeneous-Domain Knowledge Propagation [157] learns a domain translator function from multi-modal source data that can be used to predict class labels in the target even if only one of the modality is present. The proposed structure has the advantage to be flexible enough to represent both domain-specific features and shared features across domains (see Figure 22).

5 Beyond image classification

In the previous sections, we attempted to provide an overview of visual DA methods with emphasis on image categorization. Compared to this vast literature focused on object recognition, relatively few papers go beyond image classification and address domain adaptation related to other computer vision problems such as object detection, semantic segmentation, pose estimation, video event or action detection. One of the main reason is probably due to the fact that these problems are more complex and have often additional challenges and requirements (*e.g.* precision related to the localization in the case of detection, pixel level accuracy required for image segmentation, increased amount of annotation burden needed for videos, *etc.*) Moreover, adapting visual representations such as contours, deformable and articulated 2-D or 3-D models, graphs, random fields or visual dynamics, is less obvious with classical *vectorial DA* techniques.

Therefore, when these tasks are addressed in the context of domain adaptation, the problem is generally rewritten as a classification problem with vectorial feature representations and a set of predefined class labels. In this case the main challenge becomes finding the best vectorial representation for the given the

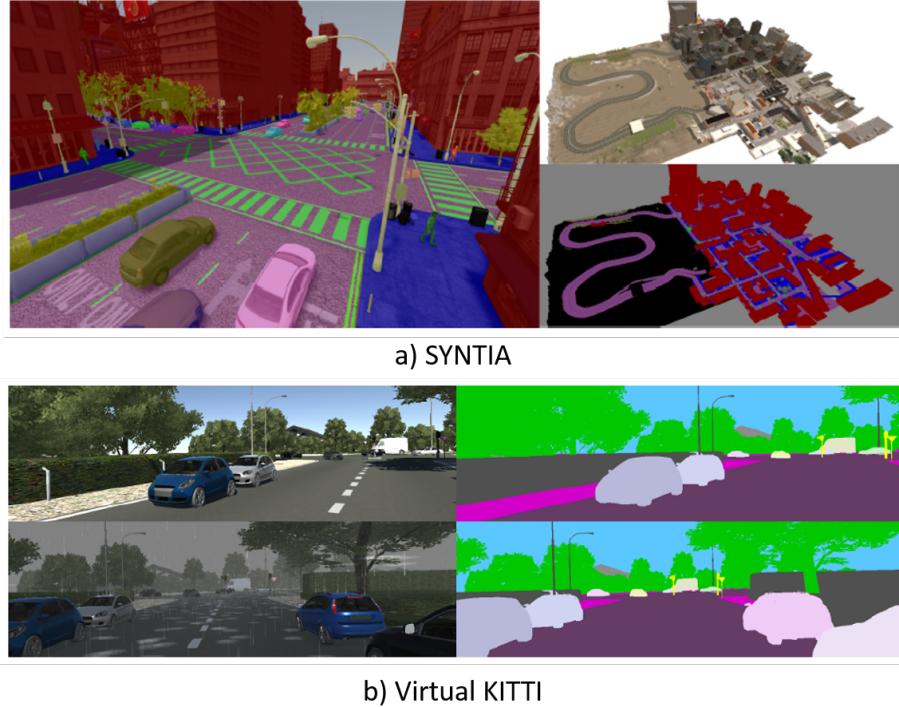


Fig. 23 Virtual word examples: SYNTHIA (top), Virtual KITTI (bottom).

task. When this is possible, shallow DA methods, described in the Section 3, can be applied to the problem. Thereupon, we can find in the literature DA solutions such as Adaptive SVM [54], DT-SVM [55], A-MKL [23] or Selective Transfer Machine [31] applied to video concept detection [22], video event recognition [23], activity recognition [24, 25], facial action unit detection [31], and 3D Pose Estimation [32].

When rewriting the problem into classification of vectorial representation is less obvious, as in the case of image segmentation, where the output is a structured output, or detection where the output is a set of bounding boxes, most often the target training set is simply augmented with the source data and traditional - segmentation, detection, *etc.* - methods are used. To overcome the lack of labels in the target domain, source data is often gathered by crawling the Web (webly supervised) [158, 159, 160] or the target set is enriched with synthetically generated data. The usage of the synthetic data became even more popular since the massive adoption of deep CNNs to perform computer vision tasks requiring large amount of annotated data.

Synthetic data based adaptation. Early methods use 3D CAD models to improve solutions for pose and viewpoint estimation [161, 162, 163, 164], object and object part detection [165, 166, 167, 168, 169, 170, 171, 172], segmentation and scene understanding [173, 174, 175]. The recent progresses in computer graphics and modern high-level generic graphics platforms such as game engines enable to generate photo-realistic

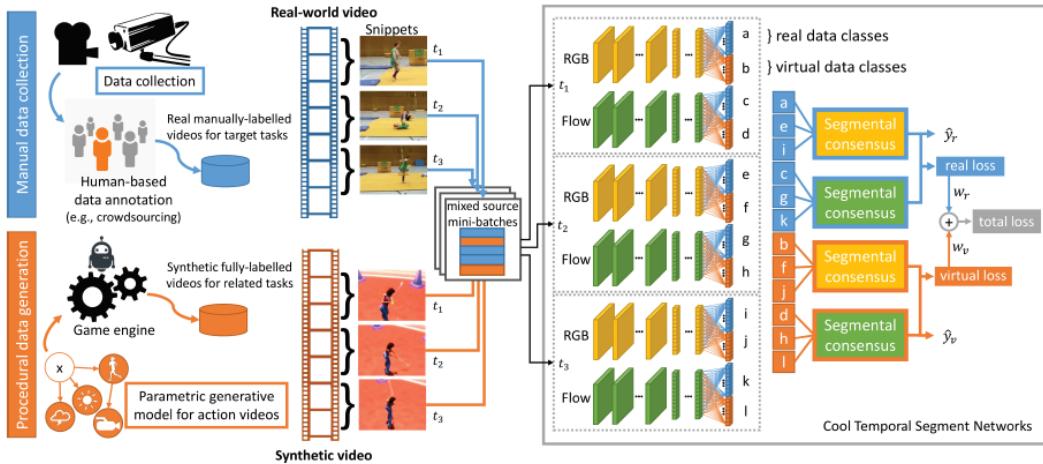


Fig. 24 Illustration of the Cool-TSN deep multi-task learning architecture [189] for end-to-end action recognition in videos. (Image courtesy C. De Souza).

virtual worlds with diverse, realistic, and physically plausible events and actions. Popular virtual worlds are SYNTHIA³⁷ [176], Virtual KITTI³⁸ [177] and GTA-V [178] (see also Figure 23).

Such virtually generated and controlled environments come with different levels of labeling for free and therefore have great promise for deep learning across a variety of computer vision problems, including optical flow [179, 180, 181, 182], object trackers [183, 177], depth estimation from RGB [184], object detection [185, 186, 187] semantic segmentation [188, 176, 178] or human actions recognition [189].

In most cases, the synthetic data is used to enrich the real data for building the models. However, DA techniques can further help to adjust the model trained with virtual data (source) to real data (target) especially when no or few labeled examples are available in the real domain [190, 191, 176, 189]. As such, [190] propose a deep spatial feature point architecture for visuomotor representation which, using synthetic examples and a few supervised examples, transfer the pretrained model to real imagery. This is done by combining a pose estimation loss, a domain confusion loss that aligns the synthetic and real domains, and a contrastive loss that aligns specific pairs in the feature space. All together, these three losses ensure that the representation is suitable to the pose estimation task while remaining robust to the synthetic-real domain shift.

The Cool Temporal Segment Network [189] is an end-to-end action recognition model for real-world target categories that combines a few examples of labeled real-world videos with a large number of procedurally generated synthetic videos. The model uses a deep multi-task representation learning architecture, able to mix synthetic and real videos even if the action categories differ between the real and synthetic sets (see Figure 24).

³⁷ Available at <http://synthia-dataset.net>

³⁸ Available at <http://www.xrce.xerox.com/Research-Development/Computer-Vision/Proxy-Virtual-Worlds>

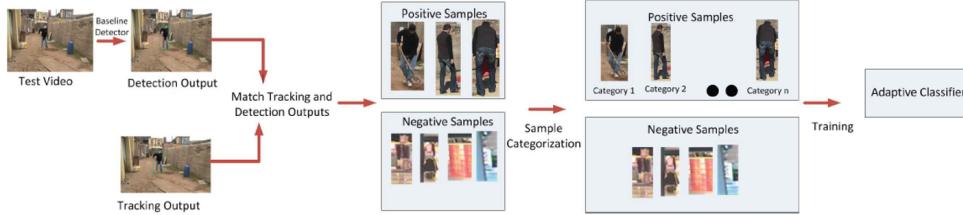


Fig. 25 Online adaptation of the generic detector with tracked regions. (Image courtesy P. Sharma [204]).

5.1 Object detection

Concerning visual applications, after the image level categorization task, object detection received the most attention from the visual DA/TL community. Object detection models, until recently, were composed of a window selection mechanism and appearance based classifiers trained on the features extracted from labeled bounding boxes. At test time, the classifier was used to decide if a region of interest obtained by sliding windows or generic window selection models [192, 193, 194] contains the object or not.

Therefore, considering the window selection mechanism as being domain independent, standard DA methods can be integrated with the appearance based classifiers to adapt to the target domain the models trained on the source domain. The Projective Model Transfer SVM (PMT-SVM) and the Deformable Adaptive SVM (DA-SVM) proposed in [195] are such methods, which adapt HOG deformable source templates [196, 197] with labeled target bounding boxes (SS scenario), and the adapted template is used at test time to detect the presence or absence of an object class in sliding windows. In [198] the PMT-SVM was further combined with MMDT [72] to handle complex domain shifts. The detector is further improved by a smoothness constraints imposed on the classifier scores utilizing instance correspondences (*e.g.* the same object observed simultaneously from multiple views or tracked between video frames).

[199] uses the TCA [14] to adapt image level HOG representation between source and target domains for object detection. [200] proposes a Taylor Expansion Based Classifier Adaptation for either boosting or logistic regression to adapt person detection between videos acquired in different meeting rooms.

Online adaptation of the detector. Most early works related to object detector adaptation concern online adaptation of a generic detector trained on strongly labeled images (bounding boxes) to detect objects (in general cars or pedestrians) in videos. These methods exploit redundancies in videos to obtain prospective positive target examples (windows) either by background modeling/subtraction [201, 202], or by combination of object tracking with regions proposed by the generic detector [203, 204, 205, 206] (see the main idea in Figure 25). Using these designated target samples in the new frame the model is updated involving semi-supervised approaches such as self-training [207, 208] or co-training [209, 210].

For instance, [211] proposes a non-parametric detector adaptation algorithm, which adjusts an offline frame-based object detector to the visual characteristic of a new video clip. The Structure-Aware Adaptive Structural SVM (SA-SSVM) [212] adapts online the deformable part-based model [213] for pedestrian detection (see Figure 26). To handle the case when no target label is available, a strategy inspired by self-paced learning and supported by a Gaussian Process Regression is used to automatically label samples in the tar-

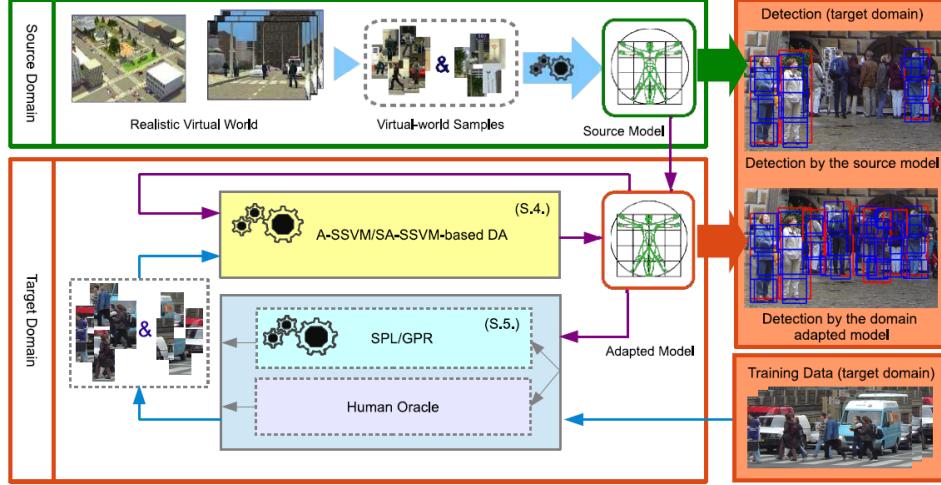


Fig. 26 Domain Adaptation of DPM based on SA-SSVM [212] (Image courtesy J. Xu).

get domains. The temporal structure of the video is exploited through similarity constraints imposed on the adapted detector.

Multi-object tracking. Multi-object tracking aims at automatically detecting and tracking individual object (*e.g.* car or pedestrian) instances [214, 205, 206]. These methods generally capitalize on multi-task and multi-instance learning to perform category-to-instance adaptation. For instance, [214] introduces a Multiple Instance Learning (MIL) loss function for Real Adaboost, which is used within a tracking based unsupervised online sample collection mechanism to incrementally adjust the pretrained detector.

[205] propose an unsupervised, online and self-tuning learning algorithm to optimize a multi-task learning based convex objective involving a high-precision/low-recall off-the-shelf generic detector. The method exploits the data structure to jointly learn an ensemble of instance-level trackers, from which adapted category-level object detectors are derived. The main idea in [206] is to jointly learn all detectors (the target instance models and the generic one) using an online adaptation via Bayesian filtering coupled with multi-task learning to efficiently share parameters and reduce drift, while gradually improving recall.

The transductive approach in [203] re-trains the detector with automatically discovered target domain examples starting with the easiest first, and iteratively re-weighting labeled source samples by scoring trajectory tracks. [204] introduces a multi-class random fern adaptive classifier where different categories of the positive samples (corresponding to different video tracks) are considered as different target classes, and all negative online samples are considered as a single negative target class. [215] proposes a particle filtering framework for multi-person tracking-by-detection to predict the target locations.

Deep neural architectures. More recently, end-to-end deep learning object detection models were proposed that integrate and learn simultaneously the region proposals and the object appearance. In general, these models are initialized by deep models pretrained with image level annotations (often on the ILSVRC datasets [122]). In fact, the pretrained deep model combined with class-agnostic region of interest proposal, can

already be used to predict the presence or absence of the target object in the proposed local regions [216, 217, 133, 218]. When strongly labeled target data is available, the model can further be fine-tuned using the labeled bounding boxes to improve both the recognition and the object localization. Thus, the Large Scale Detection through Adaptation³⁹ [218] learns to transform an image classifier into an object detector by fine-tuning the CNN model, pretrained on images, with a set of labeled bounding boxes. The advantage of this model is that it generalizes well even for localization of classes for which there were no bounding box annotations during the training phase.

Instead fine-tuning, [219] uses Subspace Alignment [19] to adjust class specific representations of bounding boxes (BB) between the source and target domain. The source BBs are extracted from the strongly annotated training set, while the target BBs are obtained with the RCNN-detector [217] trained on the source set. The detector is then re-trained with the target aligned source features and used to classify the target data projected into the target subspace.

6 Beyond domain adaptation: unifying perspectives

The aim of this section is to relate domain adaptation to other machine learning solutions. First in Section 6.1 we discuss how DA is related to other transfer learning (TL) techniques. Then, in Section 6.2 we connect DA to several classical machine learning approaches illustrating how these methods are exploited in various DA solutions. Finally, in Section 6.3 we examine the relationship between heterogeneous DA and multi-view/multi-modal learning.

6.1 DA within transfer learning

As shown in Section 2, DA is a particular case of the transductive transfer learning aimed to solve a classification task common to the source and target, by simultaneously exploiting labeled source and unlabeled target examples (see also Figure 2). As such, DA is opposite to unsupervised TL, where both domains and tasks are different with labels available neither for source nor for target.

DA is also different from self-taught learning [220], which exploits a limited labeled target data for a classification task together with a large amount of unlabeled source data mildly related to the task. The main idea behind self-taught learning is to explore the unlabeled source data and to discover repetitive patterns that could be used for the supervised learning task.

On the other hand, DA is more closely related to domain generalization [221, 222, 138, 223, 224], multi-task learning [225, 226, 227] or few-shot learning [228, 229] discussed below.

Domain generalization. Similarly to multi-source DA [83, 82, 84], domain generalization methods [221, 222, 138, 223, 224] aim to average knowledge from several related source domains, in order to learn a model for a new target domain. But, in contrast to DA where unlabeled target instances are available to adapt the model, in domain generalization, no target example is accessible at training time.

³⁹ Code available at <https://github.com/jhoffman/lstda/zipball/master>

Multi-task learning. In multi-task learning [225, 226, 227] different tasks (*e.g.* sets of the labels) are learned at the same time using a shared representation such that what is learned for each task can help in learning the other tasks. If we consider the tasks in DA as domain source and target) specific tasks, a semi-supervised DA method can be seen as a sort of two-task learning problem where, in particular, learning the source specific task helps learning the target specific task. Furthermore, in the case of multi-source domain adaptation [230, 231, 89, 87, 232, 86, 28, 233, 62, 77] different source specific tasks are jointly exploited in the interest of the target task.

On the other hand, as we have seen in Section 5.1, multi-task learning techniques can be beneficial for online DA, in particular for multi-object tracking and detection [205, 206], where the generic object detector (trained on source data) is adapted for each individual object instance.

Few-shot learning. Few-shot learning [228, 229, 89, 234] aims to learn information about object categories when only a few training images are available for training. This is done by making use of prior knowledge of related categories for which larger amount of annotated data is available. Existing solutions are the knowledge transfer through the reuse of model parameters [235], methods sharing parts or features [236] or approaches relying on contextual information [237].

An extreme case of few-shot learning is the *zero-shot learning* [238, 239], where the new task is deduced from previous tasks without using any training data for the current task. To address zero-shot learning, the methods rely either on nameable image characteristics and semantic concepts [238, 239, 240, 241], or on latent topics discovered by the system directly from the data [242, 243, 244]. In both cases, detecting these attributes can be seen as the common tasks between the training classes (source domains) and the new classes (target domains).

Unified DA and TL models. We have seen that the particularity of DA is the shared label space, in contrast to more generic TL approaches where the focus is on the task transfer between classes. However, in [245] it is claimed that task transfer and domain shift can be seen as different declinations of *learning to learn* paradigm, *i.e.* the ability to leverage prior knowledge when attempting to solve a new task. Based on this observation, a common framework is proposed to leverage source data regardless of the origin of the distribution mismatch. Considering prior models as experts, the original features are augmented with the output confidence values of the source models and target classifiers are then learned with these features.

Similarly, the Transductive Prediction Adaptation (TPA) [246] augments the target features with class predictions from source experts, before applying the MDA framework [12] on these augmented features. It is shown that MDA, exploiting the correlations between the target features and source predictions, can denoise the class predictions and improve classification accuracy. In contrast to the method in [245], TPA works also in the case when no label is available in the target domain (US scenario).

The Cross-Domain Transformation [17] learns a regularized non-linear transformation using supervised data from both domains to map source examples closer to the target ones. It is shown that the models built in this new space generalize well not only to new samples from categories used to train the transformation (DA) but also to new categories that were not present at training time (task transfer). The Unifying Multi-Domain Multi-Task Learning [247], is a Neural Network framework that can be flexibly applied to multi-task, multi-domain and zero-shot learning and even to zero-shot domain adaptation.

6.2 DA related to traditional ML methods

Semi-supervised learning. DA can be seen as a particular case of the semi-supervised learning [248, 249], where, similarly to the majority of DA approaches, unlabeled data is exploited to remedy the lack of labeled data. Hence, ignoring the domain shift, traditional semi-supervised learning can be used as a solution for DA, where the source instances form the supervised part, and the target domain provides the unlabeled data. For this reason, DA methods often exploit or extend semi-supervised learning techniques such as transductive SVM [56], self-training [207, 208, 78, 77], or co-training [209, 210]. When the domain shift is small, traditional semi-supervised methods can already bring a significant improvement over baseline methods obtained with the pretrained source model [56].

Active learning. Instance selection based DA methods exploit ideas from active learning [250] to select instances with best potentials to help the training process. Thus, the Migratory-Logit algorithm [251] explores both the target and source data to actively select unlabeled target samples to be added to the training sets. [252] describes an active learning method for relevant target data selection and labeling, which combines TrAdaBoost [49] with standard SVM. [224], (see also Chapter 15), uses active learning and DA techniques to generalize semantic object parts (*e.g.* animal eyes or legs) to unseen classes (animals). The methods described in [253, 254, 255, 78, 77, 256] combine transfer learning and domain adaptation with the target sample selection and automatic sample labeling, based on the classifier confidence. These new samples are then used to iteratively update the target models.

Online learning. Online or sequential learning [257, 258, 259] is strongly related to active learning; in both cases the model is iteratively and continuously updated using new data. However, while in active learning the data to be used for the update is actively selected, in online learning generally the new data is acquired sequentially. Domain adaptation can be combined with online learning too. As an example, we presented in Section 5.1 the online adaptation for incoming video frames of a generic object detector trained offline on labeled image sets [215, 212]. [109] proposes online adaptation of image classifier to user generated content in social computing applications.

Furthermore, as discussed in Section 4, fine-tuning a deep model [132, 128, 133, 134, 135, 21], pretrained on ImageNet (source), for a new dataset (target), can be seen as sort of semi-supervised domain adaptation. Both, fine-tuning as well as training deepDA models [147, 143, 154], use sequential learning where data batches are used to perform the stochastic gradient updates. If we assume that these batches contain the target data acquired sequentially, the model learning process can be directly used for online DA adaptation of the original model.

Metric learning. In Section 3 we presented several metric learning based DA methods [74, 17, 260, 76, 77], where class labels from both domains are exploited to bridge the relatedness between the source and target. Thus, [74] proposes a new distance metric for the target domain by using the existing distance metrics learned on the source domain. [17] uses information-theoretic metric learning [75] as a distance metric across different domains, which was extended to non-linear kernels in [76]. [77] proposes a metric learning adapted to the DSCM classifier, while [260] defines a multi-task metric learning framework to learn relationships between source and target tasks. [4] explores various metric learning approaches to align deep features extracted from RGB and NIR face images.

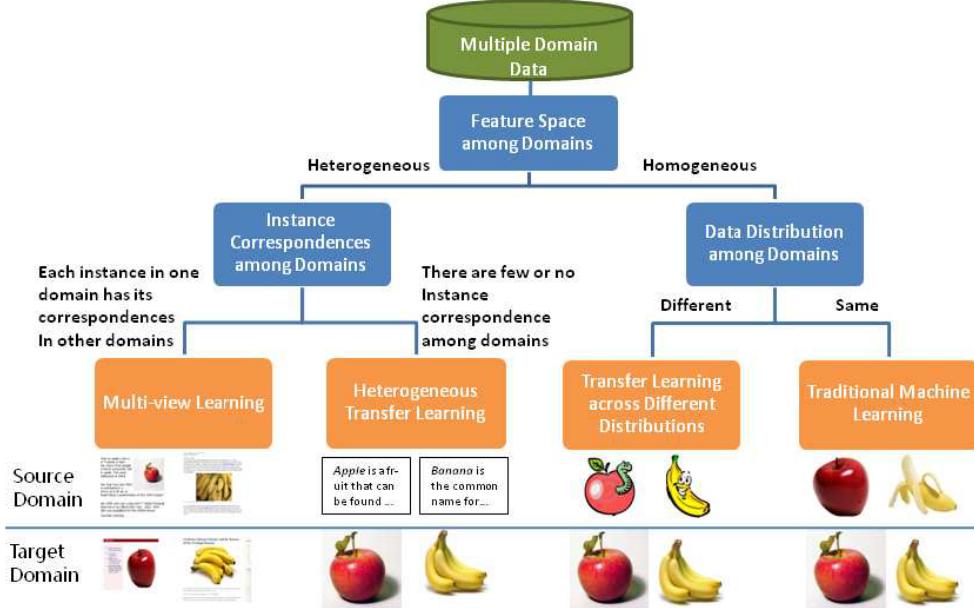


Fig. 27 Illustrating through an example the difference between TL to ML in the case of homogeneous data and between multi-view and HTL/HDA when working with heterogeneous data. Image courtesy Q. Yang [264].

Classifier ensembles. Well studied in ML, classifier ensembles have also been considered for DA and TL. As such, [261] applies a bagging approach for transferring the learning capabilities of a model to different domains where a high number of trees is learned on both source and target data in order to build a pruned version of the final ensemble to avoid a negative transfer. [262] uses random decision forests to transfer relevant features between domains. The optimization framework in [263] takes as input several classifiers learned on the source domain as well as the results of a cluster ensemble operating solely on the target domain, yielding a consensus labeling of the data in the target domain. Boosting was extended to DA and TL in [49, 51, 200, 87, 52].

6.3 HDA related to multi-view/multi-modal learning

In many data intensive applications, such as video surveillance, social computing, medical health records or environmental sciences, data collected from diverse domains or obtained from various feature extractors exhibit heterogeneity. For example, a person can be identified by different facets *e.g.* face, fingerprint, signature or iris, or in video surveillance, an action or event can be recognized using multiple cameras. When working with such heterogeneous or multi-view data most, methods try to exploit simultaneously different modalities to build better final models.

As such, multi-view learning methods are related to HDA/HTL as discussed also in Section 3.3. Nevertheless, while multi-view learning [90, 91] assumes that multi-view examples are available during training, in the case of HDA [116, 119, 84, 28, 118], this assumption rarely holds (see illustration in 27). On contrary, the aim of HDA is to transfer information from the source domain represented with one type of data (*e.g.* text) to the target domain represented with another type of data (*e.g.* images). While this assumption essentially differentiates the multi-view learning from HDA, we have seen in Section 3.3 that HDA methods often rely on an auxiliary intermediate multi-view domain [104, 106, 103, 101, 107, 109]. Hence, HDA/HTL can strongly benefit from multi-view learning techniques such as canonical correlation analysis [93], co-training [265], spectral embedding [116] and multiple kernel learning [114].

Similarly to HDA/HTL relying on intermediate domains, cross-modal image retrieval methods depend on multi-view auxiliary data to define cross-modal similarities [266, 267], or to perform semantic [268, 269, 270, 95] or multi-view embedding [93, 110, 99, 111, 112, 113]. Hence, HDA/HTL can strongly benefit from such cross-modal data representations.

In the same spirit, *webly supervised* approaches [271, 272, 273, 274, 158, 159, 275] are also related to DA and HDA as these approaches rely on collected Web data (source) data used to refine the target model. As such, [276] uses multiple kernel learning to adapt visual events learned from the Web data for video clips. [277] and [275] propose domain transfer approaches from weakly-labeled Web images for action localization and event recognition tasks.

7 Conclusion

In this chapter we tried to provide an overview of different solutions for visual domain adaptation, including both shallow and deep methods. We grouped the methods both by their similarity concerning the problem (homogeneous *vs.* heterogeneous data, unsupervised *vs.* semi-supervised scenario) and the solutions proposed (feature transformation, instance reweighing, deep models, online learning, *etc.*). We also reviewed methods that solve DA in the case of heterogeneous data as well as approaches that address computer vision problems beyond the image classification, such as object detection or multi-object tracking. Finally, we positioned domain adaptation within a larger context by linking it to other transfer learning techniques as well as to traditional machine learning approaches.

Due to the lack of the space and the large amount of methods mentioned, we could only briefly depict each method; the interested reader can follow the reference for deeper reading. We also decided not to provide any comparative experimental results between these methods for the following reasons: (1) Even if many DA methods were tested on the benchmark OFF31 [17] and OC10 [18] datasets, papers use often different experimental protocols (sampling the source *vs.* using the whole data, unsupervised *vs.* supervised) and different parameter tuning strategies (fix parameter sets, tuning on the source, cross validation or unknown). (2) Results reported in different papers given the same methods (*e.g.* GFK, TCA, SA) vary also a lot between different re-implementations. For all these reasons, making a fair comparison between all the methods based only on the literature review is rather difficult. (3) These datasets are rather small, some methods have published results only with the outdated SURFBOV features and relying only on these results is not sufficient to derive general conclusions about the methods. For a fair comparison, deep methods should be compared to shallow methods using deep features extracted from similar architectures, but both features extracted from the latest deep models and deep DA architectures build on these models perform extremely well on OFF31 and OC10 even without adaptation.

Most DA solutions in the literature are tested on these relatively small datasets (both in terms of number of classes and number of images). However, with the proliferation of sensors, large amount of heterogeneous data is collected, and hence there is a real need for solutions being able to efficiently exploit them. This shows a real need for more challenging datasets to evaluate and compare the performance of different methods. The few new DA datasets, such as the Testbed cross-dataset (TB) [278] or datasets built for model adaptation between photos, paintings and sketches [1, 2, 3, 4] while more challenging than the popular OFF31 [17], OC10 [18] or MNIST [279] vs. SVHN [280], they are only sparsely used. Moreover, except the cross-modal Place dataset [3], they are still small scale and single modality datasets.

We have also seen that only relatively few papers address adaptation beyond recognition and detection. Image and video understanding, semantic and instance level segmentation, human pose, event and action recognition, motion and 3D scene understanding, where trying to simply describe the problem with a vectorial representation and classical domain adaptation, even when it is possible, has serious limitations. Recently, these challenging problems are addressed with deep methods requiring large amount of labeled data. How to adapt these new models between domains with no or very limited amount of data is probably one of the main challenge that should be addressed by the visual domain adaptation and transfer learning community in the next few years.

References

1. B. F. Klare, S. S. Bucak, A. K. Jain, and T. Akgul, “Towards automated caricature recognition,” in *International Conference on Biometrics (ICB)*, 2012.
2. E. J. Crowley and A. Zisserman, “In search of art,” in *ECCV Workshop on Computer Vision for Art Analysis*, 2014.
3. L. Castrejón, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba, “Learning aligned cross-modal representations from weakly aligned data,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
4. S. Saxena and J. Verbeek, “Heterogeneous face recognition with cnns,” in *ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2016.
5. H. Daumé III and D. Marcu, “Domain adaptation for statistical classifiers,” *Journal of Artificial Intelligence Research*, vol. 26, no. 1, pp. 101–126, 2006.
6. S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, “Cross-domain sentiment classification via spectral feature alignment,” in *International Conference on World Wide Web (WWW)*, 2010.
7. J. Blitzer, S. Kakade, and D. P. Foster, “Domain adaptation with coupled subspaces,” in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.
8. M. Zhou and K. C. Chang, “Unifying learning to rank and domain adaptation: Enabling cross-task document scoring,” in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2014.
9. P. Prettenhofer and B. Stein, “Cross-language text classification using structural correspondence learning,” in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2010.
10. J. T. Zhou, I. W. Tsang, S. J. Pan, and M. Tan, “Heterogeneous domain adaptation for multiple classes,” in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.
11. J. T. Zhou, S. J. Pan, I. W. Tsang, and Y. Yan, “Hybrid heterogeneous transfer learning through deep learning,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2014.
12. M. Chen, Z. Xu, K. Q. Weinberger, and F. Sha, “Marginalized denoising autoencoders for domain adaptation,” in *International Conference on Machine Learning (ICML)*, 2012.
13. X. Glorot, A. Bordes, and Y. Bengio, “Domain adaptation for large-scale sentiment classification: A deep learning approach,” in *International Conference on Machine Learning (ICML)*, 2011.
14. S. J. Pan, J. T. Tsang, Ivor W. and Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *Transactions on Neural Networks*, vol. 22, no. 2, pp. 199 – 210, 2011.
15. C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models,” *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

16. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.
17. K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conference on Computer Vision (ECCV)*, 2010.
18. B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
19. B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.
20. M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.
21. B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.
22. J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.
23. L. Duan, D. Xu, and S.-F. Chang, "Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
24. N. FarajDavar, T. deCampos, D. Windridge, J. Kittler, and W. Christmas, "Domain adaptation in the context of sport video action recognition," in *BMVA British Machine Vision Conference (BMVC)*, 2012.
25. F. Zhu and L. Shao, "Enhancing action recognition by cross-domain dictionary learning," in *BMVA British Machine Vision Conference (BMVC)*, 2013.
26. H. Shen, S.-I. Yu, Y. Yang, D. Meng, and A. Hauptmann, "Unsupervised video adaptation for parsing human motion," in *European Conference on Computer Vision (ECCV)*, 2014.
27. M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.
28. S. Shekhar, V. M. Patel, H. V. Nguyen, and R. Chellappa, "Generalized domain-adaptive dictionaries," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
29. A. Sharma and D. W. Jacobs, "Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
30. B. Smith and L. Zhang, "Collaborative facial landmark localization for transferring annotations across datasets," in *European Conference on Computer Vision (ECCV)*, 2014.
31. W.-S. Chu, F. D. I. Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
32. M. Yamada, L. Sigal, and M. Raptis, "No bias left behind: Covariate shift adaptation for discriminative 3d pose estimation," in *European Conference on Computer Vision (ECCV)*, 2012.
33. B. Chidlovskii, S. Clinchant, and G. Csurka, "Domain adaptation in the absence of source domain data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2016.
34. G. Csurka, B. Chidlovskii, and S. Clinchant, "Adapted domain specific class means," in *ICCV workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2015.
35. G. Csurka, D. Larlus, A. Gordo, and J. Almazan, "What is the right way to represent document images?," *CoRR*, vol. arXiv:1603.01076, 2016.
36. K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 9, no. 3, 2016.
37. S. J. Pan and Q. Yang, "A survey on transfer learning," *Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
38. W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Self-taught clustering," in *International Conference on Machine Learning (ICML)*, 2008.
39. Z. Whang, Y. Song, and C. Zhang, "Transferred dimensionality reduction," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2008.
40. M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
41. H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, 2000.
42. B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *International Conference on Machine Learning (ICML)*, 2004.

43. M. Sugiyama, S. Nakajima, H. Kashima, P. v. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2008.
44. T. Kanamori, S. Hido, and M. Sugiyama, "Efficient direct density ratio estimation for non-stationarity adaptation and outlier detection," *Journal of Machine Learning Research*, vol. 10, pp. 1391–1445, 2009.
45. J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf, "Correcting sample selection bias by unlabeled data," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2007.
46. A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," in *Dataset Shift in Machine Learning* (J. Quiñonero Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, eds.), The MIT Press, 2009.
47. K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, pp. 49–57, 2006.
48. M. Dudík, R. E. Schapire, and S. J. Phillips, "Correcting sample selection bias in maximum entropy density estimation," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2005.
49. W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *International Conference on Machine Learning (ICML)*, 2007.
50. Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
51. S. Al-Stouhi and C. K. Reddy, "Adaptive boosting for transfer learning using dynamic updates," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2011.
52. B. Chidlovskii, G. Csurka, and S. Gangwar, "Assembling heterogeneous domain adaptation methods for image classification," in *CLEF online Working Notes*, 2014.
53. T. Joachims, "Transductive inference for text classification using support vector machines," in *International Conference on Machine Learning (ICML)*, 1999.
54. J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive SVMs," in *ACM Multimedia*, 2007.
55. L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank, "Domain transfer SVM for video concept detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
56. L. Bruzzone and M. Marconcini, "Domain adaptation problems: A dasvm classification technique and a circular validation strategy," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 32, pp. 770–787, 2010.
57. Y. Chen, G. Wang, and S. Dong, "Learning with progressive transductive support vector machine," *Pattern Recognition Letters*, vol. 24, no. 12, pp. 845–855, 2003.
58. W. Jiang, E. Zavesky, S.-F. Chang, and A. Loui, "Cross-domain learning methods for high-level visual concept classification," in *International Conference on Image Processing (ICIP)*, 2008.
59. H. Cheng, P.-N. Tan, and R. Jin, "Localized support vector machine and its efficient algorithm," in *SIAM International Conference on Data Mining (SDM)*, 2007.
60. H. Daumé III, "Frustratingly easy domain adaptation," *CoRR*, vol. arXiv:0907.1815, 2009.
61. R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: An unsupervised approach," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.
62. R. Gopalan, R. Li, and R. Chellappa, "Unsupervised adaptation across domain shifts by generating intermediate data representations," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 36, no. 11, 2014.
63. B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain invariant features for unsupervised domain adaptation," in *International Conference on Machine Learning (ICML)*, 2013.
64. J. Ni, Q. Qiu, and R. Chellappa, "Subspace interpolation via dictionary learning for unsupervised domain adaptation," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.
65. M. Bakhtashmotagh, M. Harandi, B. Lovell, and M. Salzmann, "Unsupervised domain adaptation by domain invariant projection," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.
66. M. Bakhtashmotagh, M. Harandi, B. Lovell, and M. Salzmann, "Domain adaptation on the statistical manifold," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
67. A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *Journal of Matrix Analysis and Applications*, vol. 20, no. 2, pp. 303–353, 1998.
68. M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. Yu, "Transfer sparse coding for robust image representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

69. G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia, “Semi-supervised transfer component analysis for domain adaptation in remote sensing image classification,” *Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3550–3564, 2015.
70. G. Csurka, B. Chidlovskii, S. Clinchant, and S. Michel, “Unsupervised domain adaptation with regularized domain instance denoising,” in *ECCV workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2016.
71. M. Long, J. Wang, G. Ding, S. J. Pan, and P. Yu, “Adaptation regularization: a general framework for transfer learning,” *Transactions on Knowledge and Data Engineering*, vol. 5, no. 26, pp. 1076–1089, 2014.
72. J. Hoffman, E. Rodner, J. Donahue, T. Darrell, and K. Saenko, “Efficient learning of domain-invariant image representations,” in *International Conference on Learning representations (ICLR)*, 2013.
73. E. Zhong, W. Fan, J. Peng, K. Zhang, J. Ren, D. Turaga, and O. Verscheure, “Cross domain distribution adaptation via kernel mapping,” in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2009.
74. Z.-J. Zha, t. Mei, M. Wang, Z. Wang, and X.-S. Hua, “Robust distance metric learning with auxiliary knowledge,” in *AAAI International Joint Conference on Artificial Intelligence (IJCAI)*, 2009.
75. J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning,” in *International Conference on Machine Learning (ICML)*, 2007.
76. B. Kulis, K. Saenko, and T. Darrell, “What you saw is not what you get: Domain adaptation using asymmetric kernel transforms,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
77. G. Csurka, B. Chidlovskii, and F. Perronnin, “Domain adaptation with a domain specific class means classifier,” in *ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2014.
78. T. Tommasi and B. Caputo, “Frustratingly easy NBNN domain adaptation,” in *IEEE International Conference on Computer Vision (ICCV)*, 2013.
79. N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, “Optimal transport for domain adaptation,” *CoRR*, vol. arXiv:1507.00504, 2015.
80. N. Farajidavar, T. deCampos, and J. Kittler, “Adaptive transductive transfer machines,” in *BMVA British Machine Vision Conference (BMVC)*, 2014.
81. R. Aljundi, R. Emonet, D. Muselet, and M. Sebban, “Landmarks-based kernelized subspace alignment for unsupervised domain adaptation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
82. L. Duan, D. Xu, and I. W. Tsang, “Domain adaptation from multiple sources: A domain-dependent regularization approach,” *Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 504–518, 2012.
83. R. Chattopadhyay, J. Ye, S. Panchanathan, W. Fan, and I. Davidson, “Multi-source domain adaptation and its application to early detection of fatigue,” in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2011.
84. I.-H. Jhuo, D. Liu, D. Lee, and S.-F. Chang, “Robust visual domain adaptation with low-rank reconstruction,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
85. R. Caseiro, J. F. Henriques, P. Martins, and J. Batista, “Beyond the shortest path : Unsupervised domain adaptation by sampling subspaces along the spline flow,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
86. J. Hoffman, B. Kulis, T. Darrell, and K. Saenko, “Discovering latent domains for multisource domain adaptation,” in *European Conference on Computer Vision (ECCV)*, 2012.
87. Y. Yao and G. Doretto, “Boosting for transfer learning with multiple sources,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
88. L. Ge, J. Gao, H. Ngo, K. Li, and A. Zhang, “On handling negative transfer and imbalanced distributions in multiple source transfer learning,” in *SIAM International Conference on Data Mining (SDM)*, 2013.
89. T. Tommasi and B. Caputo, “Safety in numbers: learning categories from few examples with multi model knowledge transfer,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
90. C. Xu, D. Tao, and C. Xu, “A survey on multi-view learning,” *CoRR*, vol. arXiv:1304.5634, 2013.
91. W. Wang, R. Arora, K. Livescu, and J. Bilmes, “On deep multi-view representation learning,” in *International Conference on Machine Learning (ICML)*, 2015.
92. K. Chaudhuri, S. Kakade, K. Livescu, and K. S. Sridharan, “Multi-view clustering via canonical correlation analysis,” in *International Conference on Machine Learning (ICML)*, 2009.
93. D. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neurocomputing*, vol. 16, no. 12, pp. 2639–2664, 2004.
94. R. Socher and F.-F. Li, “Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

95. F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
96. J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
97. A. Vinokourov, N. Cristianini, and J. Shawe-Taylor, "Inferring a semantic representation of text via cross-language correlation analysis," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2003.
98. M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," in *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2014.
99. A. Sharma, A. Kumar, H. Daumé III, and D. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
100. Q. Qiu, V. M. Patel, P. Turaga, and R. Chellappa, "Domain adaptive dictionary learning," in *European Conference on Computer Vision (ECCV)*, 2012.
101. b. Tan, Y. Song, E. Zhong, and Q. Yang, "Transitive transfer learning," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2015.
102. C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix tri-factorizations for clustering," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2005.
103. B. Tan, E. Zhong, M. Ng, and K. Q. Yang, "Mixed-transfer: Transfer learning over mixed graphs," in *SIAM International Conference on Data Mining (SDM)*, 2014.
104. Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang, "Heterogeneous transfer learning for image classification," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2011.
105. A. Singh, P. Singh, and G. Gordon, "Relational learning via collective matrix factorization," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2008.
106. G.-J. Qi, C. Aggarwal, and T. Huang, "Towards semantic knowledge propagation from text corpus to web images," in *International Conference on World Wide Web (WWW)*, 2011.
107. L. Yang, L. Jing, J. Yu, and M. K. Ng, "Learning transferred weights from co-occurrence data for heterogeneous transfer learning," *Transactions on Neural Networks and Learning Systems*, vol. 27, no. 11, pp. 2187–2200, 2015.
108. C. Andrieu, N. Freitas, A. Doucet, and M. Jordan, "An introduction to mcmc for machine learning," *Machine Learning*, vol. 50, no. 1, pp. 5–43, 2003.
109. Y. Yan, Q. Wu, M. Tan, and H. Min, "Online heterogeneous transfer learning by weighted offline and online classifiers," in *ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2016.
110. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *International Conference on Machine Learning (ICML)*, 2011.
111. Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2014.
112. G. Cao, A. Iosifidis, K. Chen, and M. Gabiouj, "Generalized multi-view embedding for visual recognition and cross-modal retrieval," *CoRR*, vol. arXiv:1605.09696, 2016.
113. L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
114. L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for heterogeneous domain adaptation," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 36, no. 6, pp. 1134–1148, 2012.
115. W. Li, L. Duan, D. Xu, and I. W. Tsang, "Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 36, no. 6, pp. 1134 – 1148, 2014.
116. X. Shi, Q. Liu, W. Fan, P. S. Yu, and R. Zhu, "Transfer learning on heterogeneous feature spaces via spectral transformation," in *IEEE International Conference on Data Mining (ICDM)*, 2010.
117. M. Xiao and Y. Guo, "Semi-supervised subspace co-projection for multi-class heterogeneous domain adaptation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2015.
118. T. Yao, Y. Pan, C.-W. Ngo, H. Li, and T. Mei, "Semi-supervised domain adaptation with subspace learning for visual recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
119. C. Wang and S. Mahadevan, "Heterogeneous domain adaptation using manifold alignment," in *AAAI International Joint Conference on Artificial Intelligence (IJCAI)*, 2011.
120. M. Harel and S. Mannor, "Learning from multiple outlooks," in *International Conference on Machine Learning (ICML)*, 2011.
121. D. L. Donoho, "Compressed sensing," *Transactions on Information Theory*, vol. 52, pp. 1289–1306, 2006.

122. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
123. J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International Conference on Machine Learning (ICML)*, 2014.
124. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep Convolutional Neural Networks," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.
125. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. arXiv:1409.1556, 2014.
126. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. arXiv:1512.03385, 2015.
127. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
128. S. Chopra, S. Balakrishnan, and R. Gopalan, "DLID: Deep learning for domain adaptation by interpolating between domains," in *ICML Workshop on Challenges in Representation Learning (WREPL)*, 2013.
129. Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 35, no. 8, pp. 1798–1828, 2013.
130. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.
131. E. J. Crowley and A. Zisserman, "The state of the art: Object retrieval in paintings using discriminative regions," in *BMVA British Machine Vision Conference (BMVC)*, 2014.
132. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *CoRR*, vol. arXiv:1311.2901, 2013.
133. M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
134. A. Babenko, A. Slesarev, A. Chigorin, and V. S. Lempitsky, "Neural codes for image retrieval," in *European Conference on Computer Vision (ECCV)*, 2014.
135. B. Chu, V. Madhavan, O. Beijbom, J. Hoffman, and T. Darrell, "Best practices for fine-tuning visual classifiers to new domains," in *ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2016.
136. E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *NIPS Workshop on Adversarial Training (WAT)*, 2016.
137. P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *International Conference on Machine Learning (ICML)*, 2008.
138. M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.
139. K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "Fast inference in sparse coding algorithms with applications to object recognition," *CoRR*, vol. arXiv:1010.3467, 2010.
140. R. Aljundi and T. Tuytelaars, "Lightweight unsupervised domain adaptation by convolutional filter reconstruction," in *ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2016.
141. J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993.
142. E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *CoRR*, vol. arXiv:1412.3474, 2014.
143. M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning (ICML)*, 2015.
144. B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2016.
145. M. Long, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," *CoRR*, vol. arXiv:1605.06636, 2016.
146. A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *CoRR*, vol. arXiv:1603.06432, 2016.
147. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, 2016.
148. E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

149. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.
150. M.-Y. Liu and O. Tuzel, “Coupled generative adversarial networks,” in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.
151. K. Bousmalis, N. Silberman, D. Dohan, , D. Erhan, and D. Krishnan, “Unsupervised pixel-level domain adaptation with generative adversarial networks,” *CoRR*, vol. arXiv:1612.05424, 2016.
152. D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.
153. K. Bousmalis, G. Trigeorgis, N. Silberman, D. Erhan, and D. Krishnan, “Domain separation networks,” in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.
154. M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi, “Deep reconstruction-classification networks for unsupervised domain adaptation,” in *European Conference on Computer Vision (ECCV)*, 2016.
155. M. Zeiler, D. Krishnan, G. Taylor, and R. Fergus, “Deconvolutional networks,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
156. W.-Y. Chen, T.-M. H. Hsu, and Y.-H. H. Tsai, “Transfer neural trees for heterogeneous domain adaptation,” in *European Conference on Computer Vision (ECCV)*, 2016.
157. X. Shu, G.-J. Qi, J. Tang, and W. Jingdong, “Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation,” in *ACM Multimedia*, 2015.
158. S. Divvala, A. Farhadi, and C. Guestrin, “Learning everything about anything: Webly-supervised visual concept learning,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
159. X. Chen and A. Gupta, “Webly supervised learning of convolutional networks,” in *IEEE International Conference on Computer Vision (ICCV)*, 2015.
160. E. Crowley and A. Zisserman, “The art of detection,” in *ECCV Workshop on Computer Vision for Art Analysis, (CVAA)*, 2016.
161. A. Agarwal and B. Triggs, “A local basis representation for estimating human pose from cluttered images,” in *Asian Conference on Computer Vision (ACCV)*, 2006.
162. J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-time human pose recognition in parts from single depth images,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
163. P. Panareda-Busto, J. Liebelt, and J. Gall, “Adaptation of synthetic data for coarse-to-fine viewpoint refinement,” in *BMVA British Machine Vision Conference (BMVC)*, 2015.
164. H. Su, C. Qi, Y. Yi, and L. Guibas, “Render for CNN: viewpoint estimation in images using CNNs trained with rendered 3D model views,” in *IEEE International Conference on Computer Vision (ICCV)*, 2015.
165. M. Stark, M. Goesele, and B. Schiele, “Back to the future: Learning shape models from 3D CAD data,” in *BMVA British Machine Vision Conference (BMVC)*, 2010.
166. B. Pepik, M. Stark, P. Gehler, and B. Schiele, “Teaching 3D geometry to deformable part models,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
167. B. Sun and K. Saenko, “From virtual to reality: Fast adaptation of virtual object detectors to real domains,” in *BMVA British Machine Vision Conference (BMVC)*, 2014.
168. A. Rozantsev, V. Lepetit, and P. Fua, “On rendering synthetic images for training an object detector,” *Computer Vision and Image Understanding*, vol. 137, pp. 24–37, 2015.
169. X. Peng, B. Sun, K. Ali, and K. Saenko, “Learning deep object detectors from 3D models,” in *IEEE International Conference on Computer Vision (ICCV)*, 2015.
170. H. Hattori, V. Naresh Boddeti, K. M. Kitani, and T. Kanade, “Learning scene-specific pedestrian detectors without real data,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
171. F. Massa, B. Russell, and M. Aubry, “Deep exemplar 2D-3D detection by adapting from real to rendered views,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
172. E. Bochinski, V. Eiselein, and T. Sikora, “Training a convolutional neural network for multi-class object detection using solely virtualworld data,” in *IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS)*, 2016.
173. S. Satkin, J. Lin, and M. Hebert, “Data-driven scene understanding from 3D models,” in *BMVA British Machine Vision Conference (BMVC)*, 2012.
174. L.-C. Chen, S. Fidler, and R. Yuille, Alan L. Urtasun, “Beat the MTurkers: Automatic image labeling from weak 3D supervision,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

175. J. Papon and M. Schoeler, "Semantic pose using deep networks trained on synthetic RGB-D," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.
176. G. Ros, L. Sellart, J. Materzyńska, D. Vázquez, and A. López, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
177. A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
178. S. Richter, V. Vineet, S. Roth, and K. Vladlen, "Playing for data: Ground truth from computer games," in *European Conference on Computer Vision (ECCV)*, 2016.
179. S. Meister and D. Kondermann, "Real versus realistically rendered scenes for optical flow evaluation," in *ITG Conference on Electronic Media Technology (CEMT)*, 2011.
180. D. Butler, J. Wulff, G. Stanley, and M. Black, "A naturalistic open source movie for optical flow evaluation," in *European Conference on Computer Vision (ECCV)*, 2012.
181. N. Onkarappa and A. Sappa, "Synthetic sequences and ground-truth flow field generation for algorithm validation," *Multimedia Tools and Applications*, vol. 74, no. 9, pp. 3121–3135, 2015.
182. N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
183. G. Taylor, A. Chosak, and P. Brewer, "OVVV: Using virtual worlds to design and evaluate surveillance systems," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
184. A. Shafeai, J. Little, and M. Schmidt, "Play and learn: Using video games to train computer vision models," in *BMVA British Machine Vision Conference (BMVC)*, 2016.
185. J. Marín, D. Vázquez, D. Gerónimo, and A. López, López, "Learning appearance in virtual scenarios for pedestrian detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
186. D. Vázquez, A. M. López, J. Marín, D. Ponsa, and D. Gerónimo, "Virtual and real world adaptation for pedestrian detection," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 36, no. 4, pp. 797–809, 2014.
187. J. Xu, D. Vázquez, A. López, J. Marín, and D. Ponsa, "Learning a part-based pedestrian detector in a virtual world," *Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2121–2131, 2014.
188. A. Handa, V. Patraucean, V. Badrinarayanan, S. Stent, and R. Cipolla, "Understanding real world indoor scenes with synthetic data," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
189. C. D. Souza, A. Gaidon, Y. Cabon, and A. López, "Procedural generation of videos to train deep action recognition networks," *CoRR*, vol. arXiv:1612.00881, 2016.
190. E. Tzeng, C. Devin, J. Hoffman, C. Finn, X. Peng, S. Levine, K. Saenko, and T. Darrell, "Towards adapting deep visuo-motor representations from simulated to real environments," *CoRR*, vol. arXiv:1511.07111, 2015.
191. J. Xu, S. Ramos, D. Vázquez, and A. López, "Hierarchical adaptive structural SVM for domain adaptation," *International Journal of Computer Vision*, vol. 119, no. 2, pp. 159–178, 2016.
192. X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.
193. C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision (ECCV)*, 2014.
194. J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.
195. Y. Aytar and A. Zisserman, "Tabula rasa: Model transfer for object category detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.
196. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
197. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 32, no. 9, pp. 1627–1645, 2010.
198. J. Donahue, J. Hoffman, E. Rodner, K. Saenko, and T. Darrell, "Semi-supervised domain adaptation with instance constraints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
199. F. Mirrashed, V. I. Morariu, B. Siddique, R. S. Feris, and L. S. Davis, "Domain adaptive object detection," in *Workshops on Application of Computer Vision (WACV)*, 2013.
200. C. Zhang, R. Hammid, and Z. Zhang, "Taylor expansion based classifier adaptation: Application to person detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

201. P. M. Roth, S. Sternig, H. Grabner, and H. Bischof, "Classifier grids for robust adaptive object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
202. S. Stalder, H. Grabner, and L. V. Gool, "Exploring context to learn scene specific object detectors..," in *International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2009.
203. K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller, "Shifting weights: Adapting object detectors from image to video," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.
204. P. Sharma and R. Nevatia, "Efficient detector adaptation for object detection in a video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
205. A. Gaidon, G. Zen, and J. A. Rodriguez-Serrano, "Self-learning camera: Autonomous adaptation of object detectors to unlabeled video streams..," *CoRR*, vol. arXiv:1406.4296, 2014.
206. A. Gaidon and E. Vig, "Online domain adaptation for multi-object tracking," in *BMVA British Machine Vision Conference (BMVC)*, 2015.
207. C. Rosenberg, M. Hebert, and H. Schneiderman, "Semisupervised self-training of object detection models," in *Workshops on Application of Computer Vision (WACV/MOTION)*, 2005.
208. B. Wu and R. Nevatia, "Improving part based object detection by unsupervised, online boosting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
209. O. Javed, S. Ali, and M. Shah, "Online detection and classification of moving objects using progressively improving detectors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
210. A. Levin, P. Viola, and Y. Freund, "Unsupervised improvement of visual detectors using co-training," in *IEEE International Conference on Computer Vision (ICCV)*, 2013.
211. X. Wang, G. Hua, and T. X. han, "Detection by detections: Non-parametric detector adaptation for a video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
212. J. Xu, S. Ramos, D. Vázquez, and A. López, "Domain adaptation of deformable part-based models," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 36, no. 12, pp. 2367–2380, 2014.
213. P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: a benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
214. P. Sharma, C. Huang, and R. Nevatia, "Unsupervised incremental learning for improved object detection in a video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
215. M. D. Breitenstein, F. Reichlin, E. Koller-Meier, B. Leibe, and L. Van Gool, "Online multi-person tracking-by-detection from a single, uncalibrated camera," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 31, no. 9, pp. 1820–1833, 2011.
216. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *CoRR*, vol. arXiv:1312.6229, 2013.
217. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
218. J. Hoffman, S. Guadarrama, E. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko, "LSDA: Large scale detection through adaptation," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.
219. A. Raj, V. P. N. Namboodiri, and T. Tuytelaars, "Subspace alignment based domain adaptation for rcnn detector," in *BMVA British Machine Vision Conference (BMVC)*, 2015.
220. R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng, "Self-taught learning: transfer learning from unlabeled data," in *International Conference on Machine Learning (ICML)*, 2007.
221. K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning (ICML)*, 2013.
222. Z. Xu, W. Li, L. Niu, and D. Xu, "Exploiting low-rank structure from latent domains for domain generalization," in *European Conference on Computer Vision (ECCV)*, 2014.
223. C. Gan, T. Yang, and B. Gong, "Learning attributes equals multi-source domain generalization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
224. D. Novotny, D. Larlus, and A. Vedaldi, "I have seen enough: Transferring parts across categories," in *BMVA British Machine Vision Conference (BMVC)*, 2016.
225. R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," *Machine Learning*, vol. 28, pp. 41–75, 1997.
226. T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2004.
227. B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil, "Multilinear multitask learning," in *International Conference on Machine Learning (ICML)*, 2013.

228. E. Miller, N. Matsakis, and P. Viola, "Learning from one example through shared densities on transforms," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
229. L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 28, no. 4, pp. 594–611, 2006.
230. Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation with multiple sources," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2009.
231. L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua, "Domain adaptation from multiple sources via auxiliary classifiers," in *International Conference on Machine Learning (ICML)*, 2009.
232. Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye, "A two-stage weighting framework for multi-source domain adaptation," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2011.
233. B. Gong, K. Grauman, and F. Sha, "Reshaping visual datasets for domain adaptation," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2013.
234. T. Tommasi and B. Caputo, "The more you know, the less you learn: from knowledge transfer to one-shot learning of object categories," in *BMVA British Machine Vision Conference (BMVC)*, 2009.
235. M. Fink, "Object classification from a single example utilizing class relevance pseudo-metrics," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2004.
236. E. Bart and S. Ullman, "Cross-generalization: Learning novel classes from a single example by feature replacement," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
237. K. Murphy, A. Torralba, and W. Freeman, "Using the forest to see the trees: a graphical model relating features, objects, and scenes," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2003.
238. V. Ferrari and A. Zisserman, "Learning visual attributes," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2007.
239. C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
240. M. Palatucci, D. Pomerleau, G. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2009.
241. Y. Fu, T. Hospedales, T. Xiang, Z. Fu, and S. Gong, "Transductive multi-view embedding for zero-shot recognition and annotation," in *European Conference on Computer Vision (ECCV)*, 2014.
242. V. Sharmanska, N. Quadrianto, and C. Lampert, "Augmented attribute representations," in *European Conference on Computer Vision (ECCV)*, 2012.
243. R. Layne, T. Hospedales, and S. Gong, "Re-id: Hunting attributes in the wild," in *BMVA British Machine Vision Conference (BMVC)*, 2014.
244. Y. Fu, T. Hospedales, T. Xiang, and S. Gong, "Learning multimodal latent attributes," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 36, no. 2, pp. 303–316, 2014.
245. N. Patricia and B. Caputo, "Learning to learn, from transfer learning to domain adaptation: A unifying perspective," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
246. S. Clinchant, G. Csurka, and B. Chidlovskii, "Transductive adaptation of black box predictions," in *Annual Meeting of the Association for Computational Linguistics(ACL)*, 2016.
247. Y. Yang and T. M. Hospedales, "A unified perspective on multi-domain and multi-task learning," in *International Conference on Learning representations (ICLR)*, 2015.
248. O. Chapelle, B. SchÄ¶lkopf, and A. Zien, *Semi-supervised learning*. MIT Press, 2006.
249. X. Zhu, A. Goldberg, R. Brachman, and T. Dietterich, *Introduction to semi-supervised learning*. Morgan & Claypool Publishers, 2009.
250. B. Settles, "active learning literature survey," Tech. Rep. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2010.
251. X. Liao, Y. Xue, and L. Carin, "Logistic regression with an auxiliary data source," in *International Conference on Machine Learning (ICML)*, 2005.
252. X. Shi, W. Fan, and J. Ren, "Actively transfer domain knowledge," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2008.
253. Y. Chan and H. Ng, "Domain adaptation with active learning for word sense disambiguation," in *Annual Meeting of the Association for Computational Linguistics(ACL)*, 2007.
254. P. Rai, A. Saha, H. Daumé III, and S. Venkatasubramanian, "Domain adaptation meets active learning," in *ACL Workshop on Active Learning for Natural Language Processing (ALNLP)*, 2010.
255. A. Saha, P. Rai, H. Daumé III, S. Venkatasubramanian, and S. DuVall, "Active supervised domain adaptation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2011.

256. X. Wang, T.-K. Huang, and J. Schneider, "Active transfer learning under model shift," in *International Conference on Machine Learning (ICML)*, 2014.
257. S. Shalev-Shwartz, *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, Hebrew University, 7 2007.
258. L. Bottou, *Online Algorithms and Stochastic Approximations*. Cambridge University Press, 1998.
259. S. Shalev-Shwartz, "Online learning and online convex optimization," *Foundations and Trends in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.
260. Y. Zhang and D.-Y. Yeung, "Transfer metric learning by learning task relationships," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2010.
261. T. Kamishima, M. Hamasaki, and S. Akaho, "Trbagg: A simple transfer learning method and its application to personalization in collaborative tagging," in *IEEE International Conference on Data Mining (ICDM)*, 2009.
262. E. Rodner and J. Denzler, "Learning with few examples by transferring feature relevance," in *BMVA British Machine Vision Conference (BMVC)*, 2009.
263. A. Acharya, E. Hruschka, J. Ghosh, and S. Acharyya, "Transfer learning with cluster ensembles," in *ICML Workshop on Unsupervised and Transfer Learning (WUTL)*, 2012.
264. Q. Yang, Y. Chen, G.-R. Xue, W. Dai, and Y. Yong, "Heterogeneous transfer learning for image clustering via the social-web," in *Annual Meeting of the Association for Computational Linguistics(ACL)*, 2009.
265. M. Chen, K. Q. Weinberger, and J. Blitzer, "Co-training for domain adaptation," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2011.
266. J. Ah-Pine, M. Bressan, S. Clinchant, G. Csurka, Y. Hoppenot, and J.-M. Renders, "Crossing textual and visual content in different application scenarios," *Multimedia Tools and Applications*, vol. 42, no. 1, pp. 31–56, 2009.
267. Y. Jia, M. Salzmann, and T. Darrell, "Learning cross-modality similarity for multinomial data," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.
268. J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: learning to rank with joint word-image embeddings," *Machine Learning*, vol. 81, no. 1, pp. 21–35, 2010.
269. N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM Multimedia*, 2010.
270. A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2013.
271. R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *IEEE International Conference on Computer Vision (ICCV)*, 2005.
272. X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma, "Annotating images by mining image search results," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 30, pp. 1919–1932, 2008.
273. F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," in *IEEE International Conference on Computer Vision (ICCV)*, 2007.
274. A. Bergamo and L. Torresani, "Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach," in *Annual Conference on Neural Information Processing Systems (NIPS)*, 2010.
275. C. Gan, C. Sun, L. Duan, and B. Gong, "Weby-supervised video recognition by mutually voting for relevant web images and web video frames," in *European Conference on Computer Vision (ECCV)*, 2016.
276. L. Duan, D. Xu, I. W. Tsang, and J. Luo, "Visual event recognition in videos by learning from web data," *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, vol. 34, no. 9, pp. 1667–1680, 2012.
277. C. Sun, S. Shetty, R. Sukthankar, and R. Nevatia, "Temporal localization of fine-grained actions in videos by domain transfer from web images," in *ACM Multimedia*, 2015.
278. T. Tommasi and T. Tuytelaars, "A testbed for cross-dataset analysis," in *ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV)*, 2014.
279. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
280. Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning (DLUFL)*, 2011.