Contents lists available at ScienceDirect

Neural Networks

Multi-representation adaptation network for cross-domain image classification

Yongchun Zhu [a,b], Fuzhen Zhuang [a,b,*], Jindong Wang [c], Jingwu Chen [a,b], Zhiping Shi [d], Wenjuan Wu [e], Qing He [a,b]

[a] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] Microsoft Research, China
[d] Capital Normal University, China
[e] School of Information, Renmin University of China, Beijing 100872, China

ARTICLE INFO

ABSTRACT

In image classification, it is often expensive and time-consuming to acquire sufficient labels. To solve this problem, domain adaptation often provides an attractive option given a large amount of labeled data from a similar nature but different domains. Existing approaches mainly align the distributions of representations extracted by a single structure and the representations may only contain partial information, e.g., only contain part of the saturation, brightness, and hue information. Along this line, we propose Multi-Representation Adaptation which can dramatically improve the classification accuracy for cross-domain image classification and specially aims to align the distributions of multiple representations extracted by a hybrid structure named Inception Adaptation Module (IAM). Based on this, we present Multi-Representation Adaptation Network (MRAN) to accomplish the cross-domain image classification task via multi-representation alignment which can capture the information from different aspects. In addition, we extend Maximum Mean Discrepancy (MMD) to compute the adaptation loss. Our approach can be easily implemented by extending most feed-forward models with IAM, and the network can be trained efficiently via back-propagation. Experiments conducted on three benchmark image datasets demonstrate the effectiveness of MRAN.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

As one of the fundamental technologies in computer vision, image classification has widely been researched. There are many applications of image classification, such as face recognition (Parkhi, Vedaldi, Zisserman, et al., 2015), handwritten recognition (LeCun et al., 1990), and human activity recognition (Bulling, Blanke, & Schiele, 2014). To successfully construct an image classification system, a sufficient number of manually annotated images for each specific target domain are required beforehand. With a large amount of labeled training data and substantial computation resources, the satisfying performances can be achieved by deep neural networks recently (He, Zhang, Ren, & Sun, 2016; Simonyan & Zisserman, 2015).

Nevertheless, in real situations, it is usually impractical to obtain sufficient manually labeled training data for every new scenario. Moreover, it is often prohibitively difficult and expensive to obtain enough labeled data. To alleviate this problem, domain adaptation (Pan & Yang, 2010; Zhuang, Cheng, Luo, Pan and He, 2015), which aims to adapt the feature representation learned in the source domain with rich label information to the target domain with less or even no label information, has received much attention in recent years.

Recent domain adaptation methods achieve remarkable results by embedding domain adaptation modules in the pipeline of deep feature learning to extract domain-invariant representations. This can generally be achieved by optimizing some measures of domain shift (Pan & Yang, 2010; Quionero-Candela, Sugiyama, Schwaighofer, & Lawrence, 2009), e.g., maximum mean discrepancy (Long, Cao, Wang, & Jordan, 2015; Tzeng, Hoffman, Zhang, Saenko, & Darrell, 2014), correlation distances (Sun, Feng, & Saenko, 2016; Sun & Saenko, 2016), or minimizing an approximate domain discrepancy distance through an adversarial objective with respect to a domain discriminator (Ganin & Lempitsky, 2015; Tzeng, Hoffman, Saenko, & Darrell, 2017).

Most of the recent deep domain adaptation methods are based on convolutional neural networks which have the ability

* Corresponding author at: Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China.
  E-mail addresses: zhuyongchun18s@ict.ac.cn (Y. Zhu), zhuangfuzhen@ict.ac.cn (F. Zhuang).

(a) Original Image (b) Saturation Informa- (c) Brightness Infor- (d) Hue Information
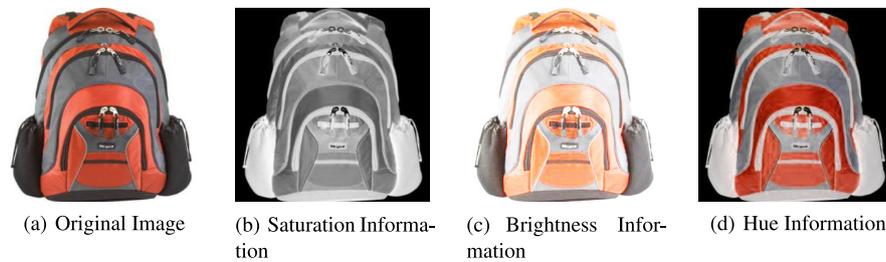tion   mation

Fig. 1. (a) is the original image, while the parts of information in (b), (c), (d) are captured from (a) by different structures. (b), (c), (d) only contain part of the saturation, brightness and hue information, respectively.
Source: The original image is from the Office31 dataset (Saenko, Kulis, Fritz, & Darrell, 2010).

to extract abstract representations from high-dimensional images. However, this feature extraction process might lose some important information. Hence, comparing to the original images, the representations may only contain partial instead of complete information, e.g., only contain part of the saturation, brightness, and hue information. An intuitive example is given in Fig. 1. Fig. 1(a) is the original image, while Figs. 1(b)∼1(d) are the transformed forms. More importantly, we find that all transformed images only contain partial information, thus they may give us wrong or distorted facts of the real image. Therefore, we need to observe the objects from multiple points to get a comprehensive understanding.

The previous deep domain adaptation methods mainly align the distributions from representations extracted from the source and target domain data by a single structure, i.e., they are single-representation adaptation methods. Similar to the transformed images in the example, the representations only contain partial information, hence the alignment also focuses on partial information. Hence, this might lead to unsatisfying transfer learning performance. To fully understand the objects, more representations should be considered when aligning the distributions. To this end, different structures of convolutional neural networks provide an option to extract multiple representations from images. Along this line, we propose Multi-Representation Adaptation (MRA), which tries to align the distributions of the source and target domains using multiple representations extracted by a hybrid neural structure.

Specifically, we propose a Multi-Representation Adaptation Networks (MRAN) to align distributions of multiple representations in a domain-specific layer across domains for unsupervised domain adaptation. To enable MRA, we propose a hybrid neural structure named Inception Adaptation Module (IAM) to extract multiple representations from images. A key novelty over previous single-representation adaptation methods is the capability of MRAN to learn multiple domain-invariant representations which contain more information. Furthermore, the nonparametric Maximum Mean Discrepancy (Gretton, Borgwardt, Rasch, Schölkopf, & Smola, 2012) (MMD) is extended to compute the adaptation loss based on conditional distribution, and integrated into deep neural networks. The IAM method can be implemented by most feed-forward models and trained efficiently using standard back-propagation. Extensive experiments performed on three benchmark datasets show that MRAN can achieve remarkable performance compared with state-of-the-art competitors.

The contributions of this paper are summarized as follows. (1) To the best of our knowledge, we are the first to learn multiple different domain-invariant representations by Inception Adaptation Module (IAM) for cross-domain image classification. (2) A novel Multi-Representation Adaptation Network (MRAN) is proposed to align distributions of multiple different representations which might contain more information about the images.

(3) MMD is extended to measure the discrepancy of conditional distributions across different domains in deep neural networks. (4) Finally, we conduct extensive experiments to validate the effectiveness of MRAN.

2. Related work

Our work mainly belongs to domain adaptation, and we will introduce the related work in three aspects: image classification, domain adaptation, and multi-view learning.

Image Classification. As one of the fundamental technologies in computer vision, image classification has widely been researched. On the basis of the assumption of the parameter on data, the image classifiers could be divided into parametric and non-parametric classifier. For parametric classifier, the parameters like mean vector and covariance matrix are frequently generated from training samples, such as Maximum likelihood, linear discriminant analysis. While non-parametric classifiers do not make use of statistical parameters to calculate class separation, such as neural network, svm, decision tree. Recently, the deep neural networks (He et al., 2016; Simonyan & Zisserman, 2015; Wang, He and Li, 2018, 2018) have achieved remarkable performance for image classification. With the guidance of the human visual system (HVS), Wang, He et al. (2018) explore the attention mechanism and propose a novel end-to-end attention recurrent convolutional network (ARCNet) for scene classification. LSLRR (Wang, He et al., 2018) improves the classical low-rank representation with locality constraint criterion and structure preserving strategy. However, they assume the training and test sets have the same distributions. Hence, these methods could not solve the cross-domain problems.

Domain Adaptation. Recent years have witnessed many approaches to solve the visual domain adaptation problem, which is also commonly framed as the visual dataset bias problem (Pan & Yang, 2010; Quionero-Candela et al., 2009; Zhuang, Luo, He and Shi, 2015). Previous shallow methods for unsupervised adaptation include re-weighting the training data so that they could more closely reflect those in the test distribution (Jiang & Zhai, 2007), and finding a transformation in a lower-dimensional manifold that draws the source and target subspaces closer (Gong, Shi, Sha, & Grauman, 2012; Pan, Tsang, Kwok, & Yang, 2011).

Most existing methods learn a shallow representation model to minimize domain discrepancy, which cannot suppress domain-specific explanatory factors of variations. Deep networks learn abstract representations that disentangle the explanatory factors of variations behind data (Bengio, Courville, & Vincent, 2013) and extract transferable factors underlying different populations (Glorot, Bordes, & Bengio, 2011; Oquab, Bottou, Laptev, & Sivic, 2014). Thus deep neural networks have been explored for domain adaptation (Ganin & Lempitsky, 2015; Long et al., 2015; Sun & Saenko, 2016; Tzeng et al., 2017, 2014; Zhu, Zhuang,

Yang, Yang and He, 2019; Zhuang, Cheng et al., 2015), where significant performance gains have been witnessed compared to prior shallow transfer learning methods.

The main strategy of deep transfer networks is to guide feature learning by minimizing the difference between the source and target distributions. Some recent works bridge deep learning and domain adaptation (Ganin & Lempitsky, 2015; Long et al., 2015; Long, Zhu, Wang, & Jordan, 2017; Pei, Cao, Long, & Wang, 2018; Tzeng, Hoffman, Darrell, & Saenko, 2015; Tzeng et al., 2017), which extend deep convolutional neural networks (CNNs) to domain adaptation. These works are mainly divided into two class, embedding methods by adding adaptation layers through which the embedding of distributions is matched (Long et al., 2015, 2017; Tzeng et al., 2014; Zhu, Zhuang and Wang, 2019), and adversarial methods by adding a subnetwork as domain discriminator while the deep features are learned to confuse the discriminator in a domain-adversarial training paradigm (Bousmalis, Silberman, Dohan, Erhan, & Krishnan, 2017; Ganin et al., 2016; Hoffman et al., 2018; Kang, Zheng, Yan, & Yang, 2018; Kumar et al., 2018; Liu & Tuzel, 2016; Saito, Watanabe, Ushiku, & Harada, 2018; Tzeng et al., 2017; Zhang, Ouyang, Li, & Xu, 2018). And recent related work extends the adversarial methods to a generative adversarial way (Bousmalis et al., 2017). Besides these two mainstreams, there are diverse methods to learn domain-invariant features: DRCN (Ghifary, Kleijn, Zhang, Balduzzi, & Li, 2016) reconstructs features to images and makes the transformed images are similar to original images. D-CORAL (Sun & Saenko, 2016) "recolors" whitened source features with the covariance of features from the target domain. All of these methods focus on aligning distributions of representations extracted by a single structure. However, the representations might only contain partial information. Our MRA could cover more information by aligning distributions of multiple representations extracted by a hybrid structure. Therefore, the representation capability can be enhanced.

Multi-view learning is concerned with the problem of machine learning from data represented by multiple distinct feature sets. The recent emergence of this learning mechanism is largely motivated by the property of data from real applications where examples are described by different feature sets or different 'views'. And multi-view learning arouses amounts of interests in the past decades (Blum & Mitchell, 1998; Gönen & Alpaydın, 2011; Yarowsky, 1995). Different from multi-view learning which needs data represented by multiple distinct feature sets, Multi-Representation learning focuses on extracting the multiple representations from the single view of data by a hybrid structure.

3. Multi-representation adaptation networks

In unsupervised domain adaptation, we are given a source domain $\mathcal{D}_s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ of n_s labeled examples where $y_i^s \in \{1, 2, \ldots, C\}$ and a target domain $\mathcal{D}_t = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$ of n_t unlabeled examples. The source domain and the target domain are sampled from different probability distributions P and Q respectively, and $P \neq Q$. The goal is to design a deep neural network $y = f(\mathbf{x})$ that formally reduces the shifts of the distributions across domains and enables learning multiple transferable representations, such that the target risk $R_t(f) = \mathbb{E}_{(\mathbf{x}, y) \, Q}[f(\mathbf{x}) \neq y]$ can be minimized by minimizing the source risk and domain discrepancy.

In recent years, the deep transfer networks have achieved remarkable results (Long et al., 2015; Tzeng et al., 2017). We call these methods as single-representation adaptation methods since they only align the distributions from representations extracted by a single structure. However, the single-representation adaptation methods focus on the partial information of the samples

mentioned above, thus they might not work well for diverse scenarios. Comparing to single-representation methods which only contain partial information, multi-representation models might cover more information. In other words, we aim to learn multiple domain-invariant representations. Along this line, we propose a hybrid structure named Inception Adaptation Module (IAM) which contains multiple substructures to extract multiple representations from images.

To achieve MRA, it is necessary to minimize the discrepancy between the distributions of the multiple representations extracted from the source and target domains. To this end, Maximum mean discrepancy (MMD) (Long et al., 2015; Tzeng et al., 2014) is extended to conditional maximum mean discrepancy (CMMD) which could compute the discrepancy of conditional distributions for multiple representations. Based on IAM and CMMD, we propose multi-representation adaptation network (MRAN). Note that, different from previous methods minimizing the discrepancy between the distributions of single representation, MRAN can align the distributions of multiple representations.

3.1. Inception adaptation module

Similar structures are adopted for recent convolutional neural networks, e.g., ResNet (He et al., 2016), DenseNet (Huang, Liu, Weinberger, & van der Maaten, 2017), and generally the structure $y = f(\mathbf{x})$ is divided into three parts $g(\cdot)$, $h(\cdot)$, $s(\cdot)$. The first part is the convolutional neural network $g(\cdot)$, which is used to convert high-pixel images to low-pixel ones; the second part $h(\cdot)$ is the global average pooling to extract representations from low-pixel images; the third part is the classifier $s(\cdot)$ to predict labels. Hence, $y = f(\mathbf{x})$ is reformulated as $y = (s \circ h \circ g)(\mathbf{x}) \left((h \circ g)(\mathbf{x}) = h(g(\mathbf{x})) \right)$.

Some recent deep transfer methods (Long et al., 2017; Pei et al., 2018) use the activations of the global average pooling layer as image representations and then align the distributions of the single representation. However, this single-representation adaptation manner might miss some important information for further performance improvement. Thus it is necessary to learn multiple domain-invariant representations by minimizing the discrepancy between the distributions of multiple representations.

To learn multiple different domain-invariant representations, the easiest way is to train multiple different convolutional neural networks. However, it would be very time-consuming to train multiple convolutional neural networks. It is well known that different structure could extract different representations from images. Hence, we use a hybrid structure IAM consisted of multiple substructures to extract multiple representations from low-pixel images. As an intuitive example shown in Fig. 2, IAM has multiple substructures $h_1(\cdot), \ldots, h_{n_r}(\cdot)$ (n_r is the number of substructures), which are different from each other. With the IAM replacing the global average pooling, multiple representations $(h_1 \circ g)(\mathbf{X}), \ldots, (h_{n_r} \circ g)(\mathbf{X})$ can be obtained. Comparing to the single representation, the multiple representations could cover more information. Hence, aligning the distributions of the multiple representations with more information could achieve better performance. The adaptation task could be achieved by minimizing the discrepancy of distributions based on the multiple representations:

$$\min_f \sum_i^{n_r} \hat{d}((h_i \circ g)(\mathbf{X}_s), (h_i \circ g)(\mathbf{X}_t)), \tag{1}$$

where $\mathbf{X}$ is the set of $\mathbf{x}$ and $\hat{d}(\cdot, \cdot)$ is an estimator of discrepancy between two distributions. To achieve the classification task, the concatenated vectors $[(h_1 \circ g)(\mathbf{X}); \ldots; (h_{n_r} \circ g)(\mathbf{X})]$ are put into the classifier $s(\cdot)$ which contains a fully connected layer and a softmax
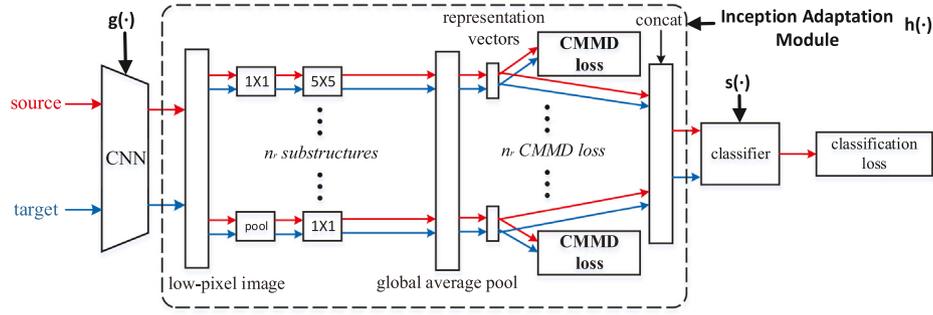
Fig. 2. Multi-Representation Adaptation Network (MRAN) aligns the conditional distributions of multiple representations. Inception Adaptation Module (IAM) could extract multiple representations from low-pixel images. By minimizing CMMD loss, the conditional distributions between the source and target domains are drawn close.

layer. The fully connected layer is mainly used to recombine the multiple representations, and the softmax layer is used to output the predicted labels. Finally, the neural network $y = f(\mathbf{x})$ with IAM is reformulated as:

$$y = f(\mathbf{x}) = s([(h_1 \circ g)(\mathbf{X}); \ldots; (h_{n_r} \circ g)(\mathbf{X})]). \quad (2)$$

Different from previous single-representation adaptation networks, the deep transfer networks with IAM are capable to learn multiple domain-invariant representations. The IAM is a multi-representation extractor. Moreover, the multiple domain-invariant representations can cover more information. It is worth noting that the IAM can be implemented by most feed-forward models. When you implement IAM in other networks, you just replace the last average pooling layer with IAM.

3.2. Conditional maximum mean discrepancy

To measure Eq. (1), a major issue is to choose a proper distance measure. We first introduce the non-parametric distance estimation Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) which has been widely used to measure the discrepancy of marginal distributions:

$$\hat{d}_{\mathcal{H}}(\mathbf{X}_s, \mathbf{X}_t) = \left\| \frac{1}{n_s} \sum_{\mathbf{x}_i \in \mathcal{D}_{\mathbf{X}^s}} \phi(\mathbf{x}_i) - \frac{1}{n_t} \sum_{\mathbf{x}_j \in \mathcal{D}_{\mathbf{X}^t}} \phi(\mathbf{x}_j) \right\|_{\mathcal{H}}^2. \quad (3)$$

By minimizing Eq. (3), the marginal distributions between the source and target domains are drawn close.

According to Elhamifar and Vidal (2013), the data samples from the same class should lay on the same subspace, even if they belong to different domains. Hence, we reduce the difference in the conditional distributions instead of marginal distributions. Indeed, minimizing the discrepancy between the conditional distributions $P_s(y_s|\mathbf{x}_s)$ and $Q_t(y_t|\mathbf{x}_t)$ is crucial for robust distribution adaptation (Sun, Chattopadhyay, Panchanathan, & Ye, 2011). Unfortunately, it is nontrivial to match the conditional distributions, even by exploring sufficient statistics of the distributions, since there are no labeled data in the target domain, i.e., $Q(y_t, \mathbf{x}_t)$ cannot be modeled directly.

Fortunately, the output of the deep neural network $\hat{y}_i^t = f(\mathbf{x}_i^t)$ could be used as the pseudo label for data in target domain. Since the posterior probabilities $P(y_s|\mathbf{x}_s)$ and $Q(y_t|\mathbf{x}_t)$ are hard to represent (Long, Wang, Ding, Sun, & Philip, 2013), we resort to explore the sufficient statistics of class-conditional distributions $P(\mathbf{x}_s|y_s = c)$ and $Q(\mathbf{x}_t|y_t = c)$ instead w.r.t. each class $c \in \{1, \ldots, C\}$. Now with the true labels of source domain data and pseudo labels of target domain data, we can essentially match the class-conditional distributions $P(\mathbf{x}_s|y_s = c)$ and $Q(\mathbf{x}_t|y_t = c)$.

Here we modify MMD to measure the distance between the class-conditional distributions $P(\mathbf{x}_s|y_s = c)$ and $Q(\mathbf{x}_t|y_t = c)$, called CMMD:

$$\hat{d}_{\mathcal{H}}(\mathbf{X}_s, \mathbf{X}_t) = \frac{1}{C} \sum_{c=1}^{C} \left\| \frac{1}{n_s^{(c)}} \sum_{\mathbf{x}_i^{s(c)} \in \mathcal{D}_{\mathbf{X}^s}^{(c)}} \phi(\mathbf{x}_i^{s(c)}) \right.$$
$$\left. - \frac{1}{n_t^{(c)}} \sum_{\mathbf{x}_j^{t(c)} \in \mathcal{D}_{\mathbf{X}^t}^{(c)}} \phi(\mathbf{x}_j^{t(c)}) \right\|_{\mathcal{H}}^2. \quad (4)$$

By minimizing Eq. (4), the conditional distributions between the source and target domains are drawn close. Though we adopt the pseudo labels of the target domain, we expect to iteratively improve the labeling quality during the optimization.

3.3. Multi-representation adaptation network

To enable effective unsupervised domain adaptation, we propose Multi-Representation Adaptation Network (MRAN) as shown in Fig. 2, which aligns the distributions of multiple representations extracted by IAM in an end-to-end deep learning model. Note that the features in the lower layers of the network are transferable and hence will not require a further distribution matching (Yosinski, Clune, Bengio, & Lipson, 2014). The loss of MRAN is formulated as:

$$\min_f \frac{1}{n_s} \sum_{i=1}^{n_s} J(f(\mathbf{x}_i^s), \mathbf{y}_i^s) + \lambda \sum_{i}^{n_r} \hat{d}((h_i \circ g)(\mathbf{X}_s), (h_i \circ g)(\mathbf{X}_t)), \quad (5)$$

where $J(\cdot, \cdot)$ is the cross-entropy loss function (classification loss), $\hat{d}(\cdot, \cdot)$ is domain adaptation loss calculated by Eq. (4), and $\lambda > 0$ is the trade-off parameter. We implement MRAN based on ResNet and replace the global average pooling by IAM. Specifically, these layers in the network are tailored to task-specific structures, which are adapted by minimizing classification error and CMMD.

Note that, training deep CNN requires a large amount of labeled data, which is prohibitive for many domain adaptation applications, so we start with the CNN pre-training on ImageNet2012 data and fine-tune it as Long et al. (2017). The training of MRAN mainly follows standard mini-batch stochastic gradient descent(SGD) algorithm. In each mini-batch, we sample the same number of source domain data and target domain data to eliminate the bias caused by domain size.

4. Experiments

We evaluate the Multi-Representation Adaptation Network (MRAN) against state-of-the-art domain adaptation methods on three datasets: **ImageCLEF-DA**, **Office-31** and **Office-Home**.

4.1. Experimental setup

4.1.1. Datasets

ImageCLEF-DA[1] is a benchmark dataset for ImageCLEF 2014 domain adaptation challenge, which is organized by selecting the 12 common categories shared by the following three public datasets, each is considered as a domain: Caltech-256 (C), ImageNet ILSVRC 2012 (I), and Pascal VOC 2012 (P). There are 50 images in each category and 600 images in each domain. We use all domain combinations and build 6 transfer tasks: I → P, P → I, I → C, C → I, C → P, P → C.

Office-31 (Saenko et al., 2010) is a benchmark for domain adaptation, comprising 4,110 images in 31 classes collected from three distinct domains: Amazon(A), which contains images downloaded from amazon.com, Webcam(W) and DSLR(D), which contain images taken by web camera and digital SLR camera with different photographical settings. The images in each domain are unbalanced across the 31 classes. To enable unbiased evaluation, we evaluate all methods on all six transfer tasks A → W, D → W, W → D, A → D, D → A, W → A as in Ganin and Lempitsky (2015), Long et al. (2017) and Tzeng et al. (2014).

Office-Home (Venkateswara, Eusebio, Chakraborty, & Panchanathan, 2017) is a new dataset which consists 15,588 images larger than Office-31 and ImageCLEF-DA. It consists of images from 4 different domains: Artistic images (A), Clip Art (C), Product images (P) and Real-World images (R). For each domain, the dataset contains images of 65 object categories collected in office and home settings.

4.1.2. Baselines

We compare MRAN with various kinds of competitors, including Transfer Component Analysis (TCA) (Pan et al., 2011), Geodesic Flow Kernel (GFK) (Gong et al., 2012), Deep Convolutional Neural Network ResNet (He et al., 2016), Deep Domain Confusion (DDC) (Tzeng et al., 2014), Deep Adaptation Network (DAN) (Long et al., 2015), Deep CORAL (D-CORAL) (Sun & Saenko, 2016), Reverse Gradient (RevGrad) (Ganin & Lempitsky, 2015), Joint Adaptation Networks (JAN) (Long et al., 2017), Multi-Adversarial Domain Adaptation (MADA) (Pei et al., 2018) and Collaborative and Adversarial Network (CAN) (Zhang et al., 2018).

To further validate the effectiveness of conditional distribution adaptation and IAM, we also evaluate several variants of MRAN: (1) MRAN (CMMD), which adds the CMMD module to ResNet; (2) MRAN (IAM), which uses IAM without adaptation loss; (3) MRAN (CMMD+IAM), which uses IAM with CMMD as the adaptation loss. Note that MRAN (CMMD) improves DAN (Long et al., 2015) by replacing the multiple MMD penalties in DAN by the CMMD penalty. Besides, MRAN (IAM) improves ResNet (He et al., 2016) by replacing the global average pooling layers by IAM. Inspired by GoogLeNet (Szegedy et al., 2015), we use four substructures ($n_r = 4$) in this work. However, you can set any number of substructures for other applications. (substructure1: conv1 × 1, conv5 × 5; substructure2: conv1 × 1, conv3 × 3, conv3 × 3; substructure3: conv1 × 1; substructure4: pool, conv1 × 1).

4.1.3. Implementation details

We employ ResNet (50 layers) to learn transferable deep representations and use the activations of the last feature layer *pool*5 as image representation for baselines (Long et al., 2017). Following standard evaluation protocols for unsupervised domain adaptation (Ganin & Lempitsky, 2015; Long et al., 2015), we use all labeled source examples as the source domain and all unlabeled target examples as the target domain. The average classification accuracy and standard error over three random

trials are reported for comparison. For all baseline methods, we either follow their original model selection procedures or conduct transfer cross-validation (Zhong, Fan, Yang, Verscheure, & Ren, 2010) if their model selection strategies are not specified. For MMD-based methods (TCA, DDC, DAN, RTN, JAN, MRAN), we adopt Gaussian kernel with bandwidth set to median pairwise squared distances on the training data (Gretton et al., 2012).

All deep methods are implemented based on the pytorch framework, and fine-tune from pytorch-provided models of ResNet (He et al., 2016). We fine-tune all convolutional and pooling layers and train the classifier layer via back propagation. Since the classifier is trained from scratch, we set its learning rate to be 10 times that of the other layers. We use mini-batch stochastic gradient descent (SGD) with momentum of 0.9 and the learning rate annealing strategy in RevGrad (Ganin & Lempitsky, 2015): the learning rate is not selected by a grid search due to high computational cost, it is adjusted during SGD using the following formula: $\eta_p = \frac{\eta_0}{(1+\alpha p)^\beta}$, where p is the training progress linearly changing from 0 to 1, $\eta_0 = 0.01$, $\alpha = 10$ and $\beta = 0.75$, which is optimized to promote convergence and low error on the source domain. To suppress noisy activations at the early stages of training, instead of fixing the adaptation factor λ, we gradually change it from 0 to 1 by a progressive schedule: $\lambda_p = \frac{2}{exp(-\gamma p)} - 1$, and $\gamma = 10$ is fixed throughout the experiments (Ganin & Lempitsky, 2015). This progressive strategy significantly stabilizes parameter sensitivity and eases model selection for MRAN.

4.2. Results

All the results of three datasets are shown in Tables 1–3, respectively. From these results, we have the following insightful observations:

- MRAN (CMMD+IAM) outperforms all comparison methods on most transfer tasks. Particularly, MRAN (CMMD+IAM) substantially improves the accuracy by large margins on ImageCLEF-DA dataset, which have the same number of images in different domains and different classes. The encouraging results indicate the importance of incorporating CMMD and IAM and validate that MRAN (CMMD+IAM) is able to learn better transferable representations.

- Comparing DAN with MRAN (CMMD) with the same Gaussian kernel, the only difference is that MRAN (CMMD) aligns the conditional distributions, while DAN aligns the marginal distributions. MRAN (CMMD) is better than DAN, and the reason may be that the data samples from the same category should lay in the same subspace, even if they belong to different domains (Elhamifar & Vidal, 2013).

- MRAN (CMMD + IAM) substantially outperforms MRAN (CMMD), which shows the importance of aligning distributions of multiple representations rather than single representation.

- MRAN (CMMD+IAM) performs better than MRAN (IAM) while other deep transfer learning methods perform better than ResNet, which indicates the importance of transfer learning.

Note that, different from all the previous deep transfer learning methods that only align the marginal distributions of representations extracted by a single structure, while our model aligns the conditional distributions of multiple representations extracted by a hybrid structure (IAM), which implies that MRAN (CMMD+IAM) has a more powerful transferable ability.

4.3. Analysis

To study the learnt representations of our model, we use MRAN (r1) as the representations extracted by the substructure1 (conv1 × 1, conv5 × 5), MRAN (r2) as the representations extracted by the substructure2 (conv1 × 1, conv3 × 3, conv3 × 3),

---

Table 1
Accuracy (%) on ImageCLEF-DA for unsupervised domain adaptation (ResNet).

Method	I → P	P → I	I → C	C → I	C → P	P → C	Avg
ResNet (He et al., 2016)	74.8 ± 0.3	83.9 ± 0.1	91.5 ± 0.3	78.0 ± 0.2	65.5 ± 0.3	91.2 ± 0.3	80.7
DDC (Tzeng et al., 2014)	74.6 ± 0.3	85.7 ± 0.8	91.1 ± 0.3	82.3 ± 0.7	68.3 ± 0.4	88.8 ± 0.2	81.8
DAN (Long et al., 2015)	75.0 ± 0.4	86.2 ± 0.2	93.3 ± 0.2	84.1 ± 0.4	69.8 ± 0.4	91.3 ± 0.4	83.3
RevGrad (Ganin & Lempitsky, 2015)	75.0 ± 0.6	86.0 ± 0.3	**96.2** ± 0.4	87.0 ± 0.5	74.3 ± 0.5	91.5 ± 0.6	85.0
D-CORAL (Sun & Saenko, 2016)	76.9 ± 0.2	88.5 ± 0.3	93.6 ± 0.3	86.8 ± 0.6	74.0 ± 0.3	91.6 ± 0.3	85.2
JAN (Long et al., 2017)	76.8 ± 0.4	88.0 ± 0.2	94.7 ± 0.2	89.5 ± 0.3	74.2 ± 0.3	91.7 ± 0.3	85.8
MADA (Pei et al., 2018)	75.0 ± 0.3	87.9 ± 0.2	96.0 ± 0.3	88.8 ± 0.3	75.2 ± 0.2	92.2 ± 0.3	85.8
CAN (Zhang et al., 2018)	78.2	87.5	94.2	89.5	75.8	89.2	85.8
MRAN (IAM)	76.2 ± 0.7	88.4 ± 0.5	91.4 ± 0.2	84.2 ± 0.1	69.2 ± 0.2	88.6 ± 0.3	83.0
MRAN (CMMD)	78.7 ± 0.2	91.1 ± 0.2	94.2 ± 0.4	88.9 ± 0.1	75.1 ± 0.3	**93.1** ± 0.1	86.9
MRAN (CMMD + IAM)	**78.8** ± 0.3	**91.7** ± 0.4	95.0 ± 0.5	**93.5** ± 0.4	**77.7** ± 0.5	**93.1** ± 0.3	**88.3**

Table 2
Accuracy (%) on Office-31 for unsupervised domain adaptation (ResNet).

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg
ResNet (He et al., 2016)	68.4 ± 0.5	96.7 ± 0.5	99.3 ± 0.1	68.9 ± 0.2	62.5 ± 0.3	60.7 ± 0.3	76.1
TCA (Pan et al., 2011)	74.7 ± 0.0	96.7 ± 0.0	99.6 ± 0.0	76.1 ± 0.0	63.7 ± 0.0	62.9 ± 0.0	79.3
GFK (Gong et al., 2012)	74.8 ± 0.0	95.0 ± 0.0	98.2 ± 0.0	76.5 ± 0.0	65.4 ± 0.0	63.0 ± 0.0	78.8
DDC (Tzeng et al., 2014)	75.8 ± 0.2	95.0 ± 0.2	98.2 ± 0.1	77.5 ± 0.3	67.4 ± 0.4	64.0 ± 0.5	79.7
DAN (Long et al., 2015)	83.8 ± 0.4	96.8 ± 0.2	99.5 ± 0.1	78.4 ± 0.2	66.7 ± 0.3	62.7 ± 0.2	81.3
D-CORAL (Sun & Saenko, 2016)	77.7 ± 0.3	97.6 ± 0.2	99.7 ± 0.1	81.1 ± 0.4	64.6 ± 0.3	64.0 ± 0.4	80.8
RevGrad (Ganin & Lempitsky, 2015)	82.0 ± 0.4	96.9 ± 0.2	99.1 ± 0.1	79.7 ± 0.4	68.2 ± 0.4	67.4 ± 0.5	82.2
JAN (Long et al., 2017)	85.4 ± 0.3	97.4 ± 0.2	99.8 ± 0.2	84.7 ± 0.3	68.6 ± 0.3	70.0 ± 0.4	84.3
MADA (Pei et al., 2018)	90.0 ± 0.1	97.4 ± 0.1	99.6 ± 0.1	**87.8** ± 0.2	**70.3** ± 0.3	66.4 ± 0.3	85.2
CAN (Zhang et al., 2018)	81.5	**98.2**	99.7	85.5	65.9	63.4	82.4
MRAN (IAM)	77.4 ± 0.5	96.1 ± 0.5	99.5 ± 0.1	81.9 ± 0.8	64.2 ± 0.4	64.8 ± 0.9	80.7
MRAN (CMMD)	87.0 ± 0.4	97.7 ± 0.2	**100.0** ± 0.0	85.8 ± 0.5	67.3 ± 0.1	66.2 ± 0.2	84.0
MRAN (CMMD + IAM)	**91.4** ± 0.1	96.9 ± 0.3	99.8 ± 0.2	86.4 ± 0.6	68.3 ± 0.5	**70.9** ± 0.6	**85.6**

Table 3
Accuracy (%) on Office-Home for unsupervised domain adaptation (ResNet).

Method	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
ResNet (He et al., 2016)	48.5	68.3	75.4	53.8	64.4	66.1	52.7	42.8	74.1	65.3	49.6	79.7	61.1
DDC (Tzeng et al., 2014)	50.5	66.5	75.0	53.6	62.6	65.1	53.2	44.8	73.7	64.1	50.8	78.2	61.5
D-CORAL (Sun & Saenko, 2016)	51.5	**68.9**	**76.3**	55.8	65.1	67.2	54.7	45.3	75.2	67.0	53.6	80.3	63.4
DAN (Long et al., 2015)	53.3	68.8	75.9	56.9	64.8	66.5	56.0	49.7	75.0	68.2	56.5	80.3	64.3
JAN (Long et al., 2017)	52.6	**68.9**	**76.3**	**57.7**	66.0	67.6	56.3	48.5	76.0	68.1	55.7	81.2	64.6
MRAN (IAM)	49.6	68.3	75.0	51.1	62.6	64.3	53.4	43.6	73.9	65.2	50.7	79.1	61.4
MRAN (CMMD)	53.5	68.7	76.1	57.5	66.1	68.2	57.3	51.4	75.9	68.3	57.4	81.1	65.1
MRAN (CMMD + IAM)	**53.8**	68.6	75.0	57.3	**68.5**	**68.3**	**58.5**	**54.6**	**77.5**	**70.4**	**60.0**	**82.2**	**66.2**

MRAN (r3) as the representations extracted by the substructure3 (conv1 × 1) and MRAN (r4) as the representations extracted by the substructure4 (pool, conv1 × 1). In addition, MRAN means the combined representations after fully connected layer.

Feature Visualization: We visualize the network representations of task A → W learned by MRAN and DAN using t-SNE embeddings (Donahue et al., 2014) in Figs. 3 and 3(f).

Comparing the representations given by MRAN (r1), MRAN (r2), MRAN (r3), MRAN (r4) in Figs. 3(b)–3(e), the multiple representations extracted by different substructures have different distributions and a different number of wrong clustering. All of these demonstrate that different neural structures have the ability to extract different representations and the multiple representations have different information. Comparing with the representations given by MRAN and DAN in Figs. 3(a) and 3(f), the combined representations given by MRAN in Fig. 3(a) show that the target categories are discriminated much more clearly.

Distribution Discrepancy: The domain adaptation theory (Ben-David et al., 2010; Mansour, Mohri, & Rostamizadeh, 2009) suggests the distribution discrepancy measure $\mathcal{A}$-distance, together with the source risk, will bound the target risk. Specifically, the proxy $\mathcal{A}$-distance is defined as $d_{\mathcal{A}} = 2(1 - 2\epsilon)$, where ϵ is the generalization error of a classifier (e.g. kernel SVM) trained on the binary problem of discriminating the source and target domain data. Fig. 4(a) shows the results of $d_{\mathcal{A}}$ on tasks A →

W, W → D with learnt representations of CNN, DAN and MRAN. The $d_{\mathcal{A}}$ of the combined representations in Fig. 4(a) are smaller than the $d_{\mathcal{A}}$ of CNN, DAN, MRAN (r1), MRAN (r2), MRAN (r3) and MRAN (r4). All of these results demonstrate that the combined representations from multiple representations are better transferable than single representation. In addition, these also prove that MRA which align distributions of multiple representations extracted by a hybrid structure could achieve better performance than previous single-representation adaptation methods.

Parameter Sensitivity: We check the sensitivity of multiple adaptation loss parameter λ, which controls the relative weight for multiple adaptation loss. Fig. 4(b) shows the performance of MRAN based on ResNet on tasks A → W and C → P by varying $\lambda \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2\}$. The accuracy of MRAN first increases and then decreases as λ varies and displays as a bell-shaped curve. From the results, we can find a proper trade-off value about 0.5 to achieve good transfer performance.

Time Complexity: CMMD loss involving pseudo label and the IAM indeed need some extra computations. We conduct additional experiments to record the time consuming of ResNet, DAN, MRAN(IAM) and MRAN(CMMD) for each iteration. All the experiments are conducted on a GeForce GTX 1080 Ti GPU, and the average time of each iteration over 100 iterations is recorded. The results are listed as follows: ResNet (0.147 s), DAN (0.277 s), MRAN(IAM) (0.173 s) and MRAN(CMMD) (0.291 s). Comparing
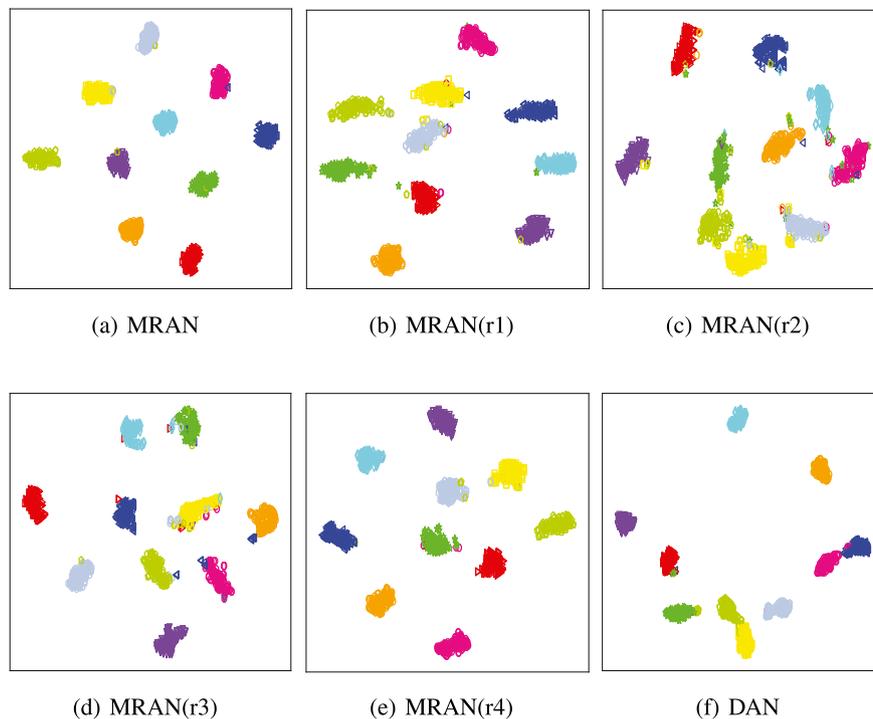
Fig. 3. Feature visualization: t-SNE of representations on source and target (a) MRAN; (b) MRAN (r1); (c) MRAN (r2); (d) MRAN (r3); (e) MRAN (r4); (f) DAN.
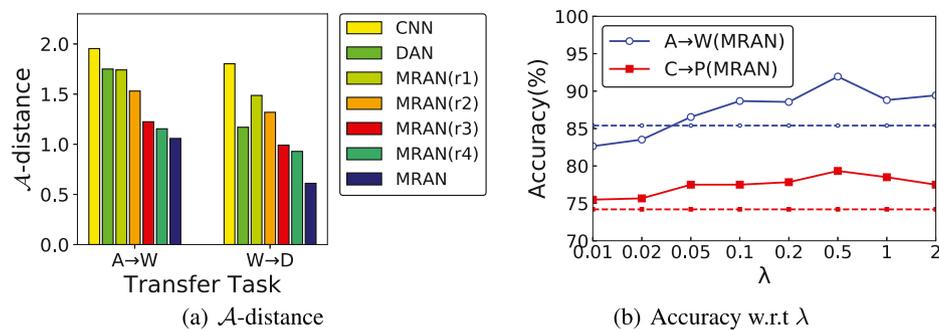

Fig. 4. (a) $\mathcal{A}$-distance; (b) parameter sensitivity of λ.

MRAN(IAM) with ResNet, we could find that the IAM takes 0.025 s more for each iteration which is very small. Comparing DAN with MRAN(CMMD), the only difference is that DAN uses MMD to align distributions and MRAN(CMMD) uses CMMD to align distributions. MRAN(CMMD) spends 0.014 s more for each iteration than DAN, which is a common property of conditional alignment methods (Pei et al., 2018). Overall, though the use of IAM and CMMD would slightly increase the time complexity, it is reasonably and can greatly improve the performance.

Insightful Findings: We get some findings from the experiments. (1) The IAM could extract multiple representations from images. The different representations could represent diverse abstract (underlying) views and we could see the t-SNE embeddings of different representations are different in Fig. 3. (2) The MRA method (our MRAN) could achieve better performance than the single-representation adaptation methods as we could see in Section 4.2, which also demonstrate the effectiveness of MRA. (3) Aligning conditional distributions are more effective than aligning marginal distributions.

5. Conclusion

To overcome the problem of aligning distributions of only one single representation from the source and target domains,

we tried to propose Multi-Representation Adaptation (MRA) to align the distributions of multiple representations. Along this line, we proposed a framework of Multi-Representation Adaptation Networks (MRAN) to learn multiple domain-invariant representations which might contain more information. In particularly, we proposed a hybrid neural structure named Inception Adaptation Module (IAM) to extract multiple representations from images. Note that, our framework can be adapted to different networks. Moreover, we extended the marginal distribution discrepancy measure MMD to conditional MMD, which is effectively incorporated in our model. Finally, extensive experiments are conducted on three datasets to demonstrate the effectiveness of the proposed model.

Acknowledgments

References

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, *79*(1–2), 151–175.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *COLT* (pp. 92–100).

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., & Krishnan, D. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR, Vol. 1* (p. 7).

Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, *46*(3), 33.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., et al. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML* (pp. 647–655).

Elhamifar, E., & Vidal, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(11), 2765–2781.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *ICML* (pp. 1180–1189).

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, *17*(1), 2096–2030.

Ghifary, M., Kleijn, W. B., Zhang, M., Balduzzi, D., & Li, W. (2016). Deep reconstruction-classification networks for unsupervised domain adaptation. In *ECCV* (pp. 597–613).

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML* (pp. 513–520).

Gönen, M., & Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, *12*(Jul), 2211–2268.

Gong, B., Shi, Y., Sha, F., & Grauman, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. In *CVPR* (pp. 2066–2073).

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*(Mar), 723–773.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).

Hoffman, J., Tzeng, E., Park, T., Zhu, J. Y., Isola, P., Saenko, K., et al. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*.

Huang, G., Liu, Z., Weinberger, K. Q., & van der Maaten, L. (2017). Densely connected convolutional networks. In *CVPR, Vol.1* (p. 3).

Jiang, J., & Zhai, C. (2007). Instance weighting for domain adaptation in nlp. In *ACL* (pp. 264–271).

Kang, G., Zheng, L., Yan, Y., & Yang, Y. (2018). Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. In *ECCV* (pp. 401–416).

Kumar, A., Sattigeri, P., Wadhawan, K., Karlinsky, L., Feris, R., Freeman, B., et al. (2018). Co-regularized alignment for unsupervised domain adaptation. In *NIPS* (pp. 9367–9378).

LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., et al. (1990). Handwritten digit recognition with a back-propagation network. In *NIPS* (pp. 396–404).

Liu, M. Y., & Tuzel, O. (2016). Coupled generative adversarial networks. *Advances in Neural Information Processing Systems*, 469–477.

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *ICML* (pp. 97–105).

Long, M., Wang, J., Ding, G., Sun, J., & Philip, S. Y. (2013). Transfer feature learning with joint distribution adaptation. In *ICCV* (pp. 2200–2207).

Long, M., Zhu, H., Wang, J., & Jordan, M. I. (2017). Deep transfer learning with joint adaptation networks. In *ICML* (pp. 2208–2217).

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009). Domain adaptation with multiple sources. In *NIPS* (pp. 1041–1048).

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR* (pp. 1717–1724).

Pan, S. J., Tsang, I. W., Kwok, J. T., & Yang, Q. (2011). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, *22*(2), 199–210.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. (2015). Deep face recognition. In *BMVC, Vol. 1* (p. 6).

Pei, Z., Cao, Z., Long, M., & Wang, J. (2018). Multi-adversarial domain adaptation. In *AAAI*.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset Shift in Machine Learning*. The MIT Press.

Saenko, K., Kulis, B., Fritz, M., & Darrell, T. (2010). Adapting visual category models to new domains. In *ECCV* (pp. 213–226).

Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR* (pp. 3723–3732).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Sun, Q., Chattopadhyay, R., Panchanathan, S., & Ye, J. (2011). A two-stage weighting framework for multi-source domain adaptation. In *NIPS* (pp. 505–513).

Sun, B., Feng, J., & Saenko, K. (2016). Return of frustratingly easy domain adaptation. In *AAAI, Vol. 6* (p. 8).

Sun, B., & Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *ECCV* (pp. 443–450).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *CVPR*.

Tzeng, E., Hoffman, J., Darrell, T., & Saenko, K. (2015). Simultaneous deep transfer across domains and tasks. In *ICCV* (pp. 4068–4076).

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *CVPR, Vol. 1* (p. 4).

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., & Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474.

Venkateswara, H., Eusebio, J., Chakraborty, S., & Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. arXiv preprint arXiv:1706.07522.

Wang, Q., He, X., & Li, X. (2018). Locality and structure regularized low rank representation for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, (99), 1–13.

Wang, Q., Liu, S., Chanussot, J., & Li, X. (2018). Scene classification with recurrent attention of vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, (99), 1–13.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *ACL* (pp. 189–196).

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. In *NIPS* (pp. 3320–3328).

Zhang, W., Ouyang, W., Li, W., & Xu, D. (2018). Collaborative and adversarial network for unsupervised domain adaptation. In *CVPR* (pp. 3801–3809).

Zhong, E., Fan, W., Yang, Q., Verscheure, O., & Ren, J. (2010). Cross validation framework to choose amongst models and datasets for transfer learning. In *ECML-PKDD* (pp. 547–562).

Zhu, Y., Zhuang, F., & Wang, D. (2019). Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *AAAI, Vol. 33* (pp. 5989–5996).

Zhu, Y., Zhuang, F., Yang, J., Yang, X., & He, Q. (2019). Adaptively transfer category-classifier for handwritten chinese character recognition. In *AKDD* (pp. 110–122). Springer.

Zhuang, F., Cheng, X., Luo, P., Pan, S. J., & He, Q. (2015). Supervised representation learning: Transfer learning with deep autoencoders. In *IJCAI* (pp. 4119–4125).

Zhuang, F., Luo, P., He, Q., & Shi, Z. (2015). Survey on transfer learning research. *Journal of Software*, *26*(1), 26–39.