Joint Adversarial Domain Adaptation

Shuang Li[1], Chi Harold Liu[1,*], Binhui Xie[1], Limin Su[1], Zhengming Ding[2], Gao Huang[3]
[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China.
[2]Department of Computer, Information and Technology, Purdue School of Engineering and Technology Indiana University-Purdue University Indianapolis.
[3]Department of Automation, Tsinghua University, Beijing, China.
{shuangli,binhuixie,sulimin}@bit.edu.cn,liuchi02@gmail.com,zd2@iu.edu,gaohuang@tsinghua.edu.cn

ABSTRACT

Domain adaptation aims to transfer the enriched label knowledge from large amounts of source data to unlabeled target data. It has raised significant interest in multimedia analysis. Existing researches mainly focus on learning domain-wise transferable representations via statistical moment matching or adversarial adaptation techniques, while ignoring the class-wise mismatch across domains, resulting in inaccurate distribution alignment. To address this issue, we propose a *Joint Adversarial Domain Adaptation (JADA)* approach to simultaneously align domain-wise and class-wise distributions across source and target in a unified adversarial learning process. Specifically, JADA attempts to solve two complementary minimax problems jointly. The feature generator aims to not only fool the well-trained domain discriminator to learn domain-invariant features, but also minimize the disagreement between two distinct task-specific classifiers' predictions to synthesize target features near the support of source class-wisely. As a result, the learned transferable features will be equipped with more discriminative structures, and effectively avoid mode collapse. Additionally, JADA enables an efficient end-to-end training manner via a simple back-propagation scheme. Extensive experiments on several real-world cross-domain benchmarks, including VisDA-2017, ImageCLEF, Office-31 and digits, verify that JADA can gain remarkable improvements over other state-of-the-art deep domain adaptation approaches.

CCS CONCEPTS

• **Computing methodologies** → **Transfer learning**; *Image representations*; • **Computer systems organization** → *Neural networks*.

KEYWORDS

Domain Adaptation; Adversarial Learning; Back-propagation; Gradient Reversal Layer; Representation Learning

*Corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351070

ACM Reference Format:
Shuang Li[1], Chi Harold Liu[1,*], Binhui Xie[1], Limin Su[1], Zhengming Ding[2], Gao Huang[3]. 2019. Joint Adversarial Domain Adaptation. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351070

1 INTRODUCTION

The past decades have witnessed a resurgence of deep convolutional neural networks [17], which has led to tremendous advances across a wide range of multimedia tasks, such as media interpretation [8, 47], multimodal retrieval [22, 50] and so on. However, in spite of the excellent learning capacity, the impressive performance of deep networks largely relies on massive amounts of well-labeled training data. Unfortunately, manually annotating sufficient label information for various multimedia applications is always time-consuming and expensive. Moreover, the well-trained deep models may perform poorly when adapting to new datasets or tasks due to the issues of *dataset bias* or *domain shift* [1, 2, 44]. Hence, there is a strong motivation to leverage the enriched knowledge of a well-annotated domain (i.e., *source domain*) to facilitate learning effective models for a different label-scarce domain (i.e., *target domain*), which is commonly referred to as domain adaptation (DA).

Generally, reducing the distribution discrepancy across domains is of vital importance to address DA problems. To this end, previous shallow DA methods are mainly based on either reweighting samples [15, 20] or learning domain-invariant features [6, 21, 32]. Recent works [7, 46] indicate that deep neural networks are able to learn more transferable representations by disentangling the explanatory factors behind source and target data. Particularly, deep CNNs are capable of obtaining general low-level features across domains to some extent. But [51] reveals that deep features can only reduce, but not remove, the cross-domain discrepancy. Hence, deep DA approaches are explored to integrate domain adaptation into deep learning pipelines by minimizing some statistical metrics [29, 36, 53], such as maximum mean discrepancy [12], or exploring adversarial learning techniques [10, 40, 45]. Most of them are devoted to bridging the cross-domain gap by explicitly aligning domain-wise distributions. Once accomplished, the model derived in the source domain can be applied to the target domain directly.

Although minimizing the domain-wise difference across domains can assist in learning deep domain-invariant representations, it may mix up the discriminative structures inevitably. As a result, this strategy will lead to negative alignment of the corresponding classes across domains, and the target samples near class decision boundaries would be easily predicted wrongly. Therefore, to alleviate this problem, we should align source and target domains with class

discriminative information preserved. Besides, without considering complex discriminative structures underlying the data, many adversarial learning based deep DA approaches [10, 45] may eventually result in mode collapse if they are not optimized appropriately.

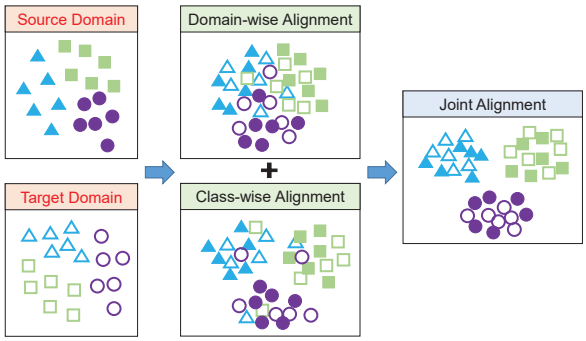

Figure 1: Illustration of our proposed JADA, which attempts to conduct joint alignment across domains based on a unified adversarial learning process, by alleviating the discriminative information loss and class-wise mismatching issues caused by domain-wise and class-wise alignment approaches respectively.

Aiming to fully leverage the discriminative structures behind source and target, Saito et al. in [39] propose a class-wise alignment approach by taking into consideration the task-specific decision boundaries. They exploit two task-specific classifiers as a discriminator to detect target samples outside the support of the source, and facilitate generating target features near source class-wisely. Since class-wise alignment approaches heavily rely on the precision of source classifier, their performance will degrade dramatically once the domain discrepancy is tremendously large, which is common in real-world multimedia tasks. Thus, jointly mitigating the domain-wise mismatching can further facilitate the precise class-wise alignment across source and target.

In this paper, to overcome the aforementioned challenges, we propose a novel *Joint Adversarial Domain Adaptation (JADA)* approach based on adversarial learning shown in Figure 1. In JADA, we jointly conduct *domain-wise* alignment by using a domain discriminator and *class-wise* alignment by exploring the task-specific label predictors, since both domain-level knowledge and discriminative information are critical to secure successful domain adaptation. To be specific, in our architecture, there are three players: a domain discriminator, a discrepancy discriminator consisting of two task-specific label predictors and a feature generator, corresponding to two minimax games. Both the domain discriminator and discrepancy discriminator acquire features from the feature generator. The domain discriminator tries to distinguish the features from source or target domains, which aims to minimize the domain prediction loss. Meanwhile, the discrepancy discriminator manages to correctly predict source samples, and specifically detects the ambiguous target samples by maximizing the output discrepancy of two label predictors. In contrast, the feature generator attempts to simultaneously cheat the domain and discrepancy discriminators. Therefore, the feature generator could yield domain-invariant feature representations with discriminative structures preserved. As a result, the learned domain-invariant features can transfer the domain-level knowledge from source to target effectively, and the exploration of discriminative structures within the data will further facilitate matching the class-wise distributions, which leads to a more precise classification performance. It is worth mentioning that, in JADA, all the training processes can be achieved in an adversarial manner by introducing an efficient gradient reversal layer (GRL) as [10] with the gradients computed by back-propagation. To sum up, we have three-fold contributions as follows:

- A novel DA adversarial learning architecture is proposed by jointly optimizing two minimax games with a domain discriminator, a discrepancy discriminator and a feature generator involved. The learning process can be achieved in an efficient end-to-end training manner via back-propagation.
- We incorporate the domain-wise and class-wise alignments into our approach, which can be verified that they are complementary to each other for DA problems. As a result, we can alleviate the mode collapse issues and improve the transfer recognition performance significantly.
- Comprehensive experiments on various cross-domain benchmarks show that JADA outperforms the state-of-the-art methods by a large margin. Moreover, analytical experiments are conducted to further verify the effectiveness of our approach, and the promising power towards real-world multimedia applications.

2 RELATED WORK

Deep domain adaptation methods concentrate on mitigating the domain discrepancy between domains by means of powerful deep neural networks [4, 10, 18, 24, 27, 29, 41, 46, 48]. Existing methods can be roughly divided into two major categories: discrepancy-based and adversarial-based methods [49].

2.1 Discrepancy-based Methods

Discrepancy-based methods mainly aim to reduce the domain shift by minimizing some discrepancy metrics, such as maximum mean discrepancy (MMD) [12, 27, 46], central moment discrepancy (CMD) [52], correlation alignment (CORAL) loss [5, 42, 43] and so on.

To name a few, Tzeng et al. [46] introduced a deep domain confusion (DDC) network to minimize the MMD distance of source and target representations at the last fully-connected layer. Further, domain adaptation network (DAN) [27] extended DDC to incorporate multiple kernel MMD distances across domains among the last three task-specific layers, which achieved better performance. To align the joint distributions of multiple domain-specific layers, Long et al. presented a joint adaptation network (JAN) as well as a joint MMD criterion [29]. Different from these MMD-based methods, Sun et al. [43] proposed a Deep CORAL method to align correlations of layer activations in deep networks. Based on deep CORAL, [53] also minimized the second-order correlation statistics of the attention maps across domains. Chen et al. [5] conducted joint domain alignment and discriminative feature learning by adding instance-based and center-based loss terms to the classic correlation alignment loss.

However, these methods would have limitations to deal with complex multimodal distribution alignment without explicitly exploring the discriminative structures underlying data distributions.

JADA could alleviate this issue and obtain fine-grained alignment across domains via a unified adversarial learning process.

2.2 Adversarial-based Methods

Representative methods in this category focus on matching feature distributions via adversarial learning, inspired by the idea of generative adversarial networks [11], which has become an increasing popular idea for addressing DA problems through an trainable adversarial manner.

Specifically, the domain-adversarial neural networks (DANN) [10] first leveraged adversarial learning between the domain classifier and feature generator to learn domain-invariant representations by adding a simple gradient reversal layer (GRL). Further, to address mode collapse issue, [34] presented a multi-adversarial domain adaptation (MADA) approach by utilizing multiple domain classifiers. Unlike these two methods, Tzeng et al. [45] summarized a general adversarial adaptation framework, then proposed an adversarial discriminative domain adaptation (ADDA) with a GAN-based loss. In [26], coupled generative adversarial networks (CoGAN) aimed to learn a joint distribution of multi-domain images by training two GANs. Based on CoGAN, [25] developed a general framework for unsupervised image-to-image translation (UNIT), and applied UNIT to DA problems by adapting the trained source classifier to classify target unlabeled samples. By considering pixel-level and low-level domain alignment, cycle-consistent adversarial domain adaptation (CyCADA) proposed in [14] conducted both generative image space and latent representation space alignment. Pixel-level domain adaptation (PixelDA) [3] tried to learn a transformation in the pixel space from source to target. [40] introduced a hierarchical adversarial deep network (HADN) to optimize the feature-level and pixel-level features by utilizing a hierarchical network structure. By contrast, a key improvement of JADA over existing adversarial learning based DA methods is the capability to jointly capture domain-level valuable knowledge and discriminative structures, which facilitates achieving satisfactory performance when the domain gap is large.

The most related idea to ours is Maximum Classifier Discrepancy (MCD) proposed in [39], which leverages two distinct task-specific decision boundaries, instead of the classical domain classifier, to align source and target class-wisely. However, since MCD only focuses on class-wise alignment, the performance will drop significantly if lots of target samples are misclassified by the two source classifiers simultaneously. JADA could effectively address this problem by matching domain-wise distributions jointly, which will improve the precision of task-specific source classifiers tremendously. Moreover, compared to MCD, JADA could optimize networks via a simple back-propagation strategy efficiently.

3 JOINT ADVERSARIAL DOMAIN ADAPTATION

Adversarial learning can be embedded into deep networks to effectively learn transferable features across source and target domains. Most existing adversarial DA methods explore either domain-wise [10] or class-wise alignment [34, 39] to mitigate the domain discrepancy. However, if the difference between domains is tremendously large, which is very common in multimedia analysis, only conducting domain-wise alignment may destroy the category discriminativeness of learned features. While only deploying class-wise alignment may be restricted by the mismatched categories without exploiting global domain knowledge. In such scenario, the cross-domain prediction accuracy will degrade dramatically. Hence, to jointly match domain-wise and class-wise distribution differences, we design a novel architecture JADA to play two minimax games in one deep neural network shown in Figure 2, which could capture global information as well as category-wise intrinsic information to enhance distribution matching between source and target domains.

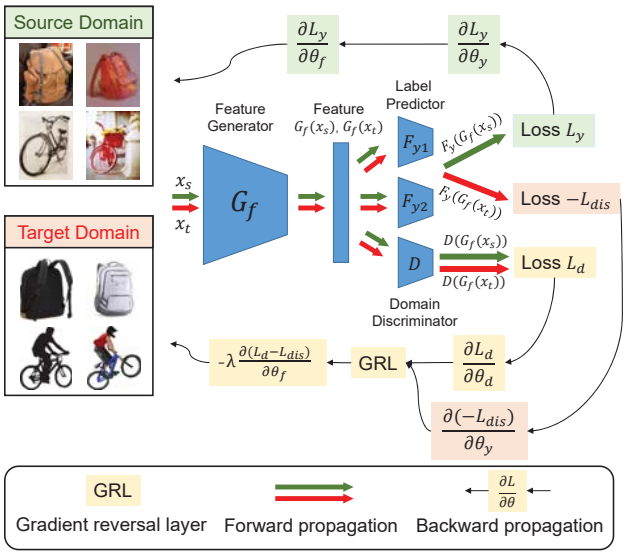

Figure 2: (Best viewed in color.) The architecture overview of JADA, where G_f and D are the feature generator and domain discriminator; $G_f(\boldsymbol{x}_s)$ and $G_f(\boldsymbol{x}_t)$ are learned features for source and target data; F_{y_1} and F_{y_2} are two task-specific label predictors; L_y, $-L_{dis}$, and L_d are the losses for label predictions, discrepancy discriminator and domain discriminator, respectively; GRL stands for a gradient reversal layer as [10]; red and green arrows represent the data flows of source and target data.

3.1 Preliminaries and Motivation

In DA problem, we denote the source domain $\mathcal{D}_s = \{(\boldsymbol{x}_{si}, y_{si})\}_{i=1}^{n_s}$ with n_s labeled samples, and the target domain $\mathcal{D}_t = \{(\boldsymbol{x}_{tj})\}_{j=1}^{n_t}$ with n_t unlabeled samples, where y_s is the corresponding label for $\boldsymbol{x}_s$. $n = n_s + n_t$ indicates the number of all the source and target samples. Since the two domains' distributions are different, the goal of our method is to design a novel deep adversarial network that enables learning features $\boldsymbol{f} = G_f(\boldsymbol{x})$ indistinguishable domain-wisely as well as class-wisely, where G_f is the feature generator in JADA.

Actually, there exist three technical challenges in adversarial-based DA: (1) capturing the global domain knowledge for matching source and target domains; (2) exploring category discriminative structures underlying distributions for accurate class-wise alignment; (3) avoiding mode collapse issues encountered in adversarial

learning. The challenges motivate our joint adversarial domain adaptation approach. To be specific, in JADA, the adversarial learning between the feature generator and the domain discriminator ensures the effective domain-wise alignment. Further, the adversarial learning between the feature generator and the discrepancy discriminator leads to generate target features near the support of source class-wisely. JADA elaborately incorporates these two adversarial learning processes into one deep architecture via optimizing a three-player game, and enables an efficient end-to-end training using classical back-propagation technique.

3.2 Domain-wise Adversarial Learning

Adversarial learning [10, 34] has been successfully applied to various DA problems, which plays a two-player game to learn domain-wise invariant features. As shown in Figure 2, the domain discriminator D aims to distinguish source and target samples accurately. On the contrary, the feature extractor G_f is trained with the purpose of confusing the domain discriminator. To reduce the distribution shift in the shared feature space, the parameters θ_d of D are optimized by minimizing the loss of domain discriminator while the parameters θ_f of G_f are trained to maximize the domain prediction loss. Here, we define $d_i = 1$ if $\boldsymbol{x}_i \in \mathcal{D}_s$, and $d_i = 0$ if $\boldsymbol{x}_i \in \mathcal{D}_t$. In addition, to guarantee the effectiveness of source model, the loss of the label predictors F_{y_1}, F_{y_2} on the annotated source data is also minimized, where θ_{y_1}, θ_{y_2} are their corresponding parameters. It is worth noting that, different from the classical DANN [10], there exist two task-specific label predictors in JADA, which aim to match class-wise distributions across domains via an adversarial manner. Therefore, the loss function of domain-wise alignment is formulated as:

$$\mathcal{J}_0(\theta_f, \theta_{y_1}, \theta_{y_2}, \theta_d) = \frac{1}{2n_s} \sum_{j=1}^{2} \sum_{i=1}^{n_s} L_y(F_{y_j}(G_f(\boldsymbol{x}_{si})), y_{si}) - \frac{\lambda}{n} \sum_{\boldsymbol{x}_i \in (\mathcal{D}_s \cup \mathcal{D}_t)} L_d(G_d(G_f(\boldsymbol{x}_i)), d_i), \quad (1)$$

where λ is a trade-off parameter to balance the label prediction loss $L_y(\cdot, \cdot)$ and the domain classification loss $L_d(\cdot, \cdot)$. In this paper, the cross-entropy loss is applied.

To extract global domain-invariant features, we are seeking the optimal parameters of $\widehat{\theta}_f$, $\widehat{\theta}_{y_1}$, $\widehat{\theta}_{y_2}$ and $\widehat{\theta}_d$, which will deliver a saddle point of Eq. (1) after the training convergence:

$$(\widehat{\theta}_d) = \arg\max_{\theta_d} \mathcal{J}_0(\theta_f, \theta_{y_1}, \theta_{y_2}, \theta_d), \quad (2)$$

$$(\widehat{\theta}_f, \widehat{\theta}_{y_1}, \widehat{\theta}_{y_2}) = \arg\min_{\theta_f, \theta_{y_1}, \theta_{y_2}} \mathcal{J}_0(\theta_f, \theta_{y_1}, \theta_{y_2}, \theta_d). \quad (3)$$

In this way, the learned features $\boldsymbol{f}$ will be indistinguishable between domains and capture the global domain knowledge, which has been proven to be powerful for deep DA problems.

3.3 Class-wise Adversarial Learning

In practical multimedia DA problems, only conducting domain-wise alignment may mix up source and target samples in the learned space and reduce the category discriminativeness to some extent, especially for the data that are near the task-specific decision boundaries. Thus, in order to match source and target distributions with discriminative information preserved, we notice that the classification boundary given by the source classifier provides strong signals to reveal the category structure information. Inspired by [39], we employ two source classifiers to detect target samples that are near the decision boundaries, which means these target samples are likely to be misclassified by the class boundaries. JADA aims to enforce the feature extractor G_f to generate target features near the support of source, which could facilitate matching domains class-wisely.

Consider two source label predictors F_{y_1} and F_{y_2} with different characteristics in Figure 2. They can be initialized differently to obtain two distinct source classifiers, which can classify source data correctly. If we use F_{y_1}, F_{y_2} to predict the labels of target data, these near decision boundaries' target samples are likely to be predicted differently by F_{y_1} and F_{y_2} from an intuitive respect. Hence, in the class-wise adversarial learning, we enforce the disagreement of F_{y_1} and F_{y_2} on the predictions for target samples to be as large as possible, which will help detect the target samples that are outside the support of source. On the contrary, the role of feature generator G_f is to minimize this disagreement to align source and target class-wisely. The game process between the feature generator and two distinct label predictors can achieve an equilibrium point via an adversarial manner.

Specifically, we denote F_{y_1} and F_{y_2} to construct a discrepancy discriminator, and the L1-norm of the disagreement between their predictions of target sample $\boldsymbol{x}_t$ as discrepancy loss:

$$L_{dis}(\boldsymbol{p}_t^{y_1}, \boldsymbol{p}_t^{y_2}) = \| \boldsymbol{p}_t^{y_1} - \boldsymbol{p}_t^{y_2} \|_1, \quad (4)$$

where $\boldsymbol{p}_t^{y_1}$ and $\boldsymbol{p}_t^{y_2}$ are probabilistic outputs of F_{y_1} and F_{y_2} for $\boldsymbol{x}_t$, respectively.

Therefore, we formulate the loss function of class-wise adversarial learning as:

$$\mathcal{J}_1(\theta_f, \theta_{y_1}, \theta_{y_2}) = \frac{1}{n_t} \sum_{i=1}^{n_t} L_{dis}(F_{y_1}(G_f(\boldsymbol{x}_{ti})), F_{y_2}(G_f(\boldsymbol{x}_{ti}))). \quad (5)$$

To achieve the class-wise adversarial goals, we want the optimal parameters $\widehat{\theta}_f, \widehat{\theta}_{y_1}$ and $\widehat{\theta}_{y_2}$ to jointly satisfy

$$(\widehat{\theta}_{y_1}, \widehat{\theta}_{y_2}) = \arg\max_{\theta_{y_1}, \theta_{y_2}} \mathcal{J}_1(\theta_f, \theta_{y_1}, \theta_{y_2}), \quad (6)$$

$$(\widehat{\theta}_f) = \arg\min_{\theta_f} \mathcal{J}_1(\theta_f, \theta_{y_1}, \theta_{y_2}). \quad (7)$$

3.4 Overall Formulation and Optimization

In this study, we simultaneously optimize two minimax problems by considering both domain-wise and class-wise distributions alignment via a straightforward back-propagation way [10]. To summarize the previous discussions, we obtain the overall objective function of JADA as:

$$\mathcal{J}(\theta_f, \theta_{y_1}, \theta_{y_2}, \theta_d) = \frac{1}{2n_s} \sum_{j=1}^{2} \sum_{i=1}^{n_s} L_y(F_{y_j}(G_f(\boldsymbol{x}_{si})), y_{si}) - \frac{\lambda}{n} \sum_{\boldsymbol{x}_i \in (\mathcal{D}_s \cup \mathcal{D}_t)} L_d(G_d(G_f(\boldsymbol{x}_i)), d_i) + \frac{\lambda}{n_t} \sum_{i=1}^{n_t} L_{dis}(F_{y_1}(G_f(\boldsymbol{x}_{ti})), F_{y_2}(G_f(\boldsymbol{x}_{ti}))), \quad (8)$$

Actually, L_y is the supervised loss on labeled source data, and L_d, L_{dis} denote the losses for domain-wise and class-wise distributions

alignment respectively. If we down-weigh the coefficient of L_d, the target prediction will drop when the domain discrepancy is large. While if we down-weigh the coefficient of L_{dis}, the category discriminative knowledge cannot be preserved, which leads to the target prediction accuracies decrease.

It is worth noting that, the optimization of Eq. (8) is not a simple combination of Eq. (1) and Eq. (5), since they will derive a contradict optimization result. To address this problem, we elaborately optimize Eq. (8) with respect to the data and gradient flows of source and target domains as:

$$(\widehat{\theta}_d) = \arg\max_{\theta_d} \mathcal{J}(\theta_f, \theta_{y_1}, \theta_{y_2}, \theta_d; \boldsymbol{x}_s, \boldsymbol{x}_t), \tag{9}$$

$$(\widehat{\theta}_f) = \arg\min_{\theta_f} \mathcal{J}(\theta_f, \theta_{y_1}, \theta_{y_2}, \theta_d; \boldsymbol{x}_s, \boldsymbol{x}_t), \tag{10}$$

$$(\widehat{\theta}_{y_1}, \widehat{\theta}_{y_2}) = \arg\min_{\theta_{y_1}, \theta_{y_2}} \mathcal{J}(\theta_f, \theta_{y_1}, \theta_{y_2}, \theta_d; \boldsymbol{x}_s), \tag{11}$$

$$(\widehat{\theta}_{y_1}, \widehat{\theta}_{y_2}) = \arg\max_{\theta_{y_1}, \theta_{y_2}} \mathcal{J}(\theta_f, \theta_{y_1}, \theta_{y_2}, \theta_d; \boldsymbol{x}_t). \tag{12}$$

Remarks: The goal of Eq. (9) is to differentiate the source and target data by minimizing the domain discriminator loss. Contrarily, Eq. (10) aims to maximize the domain discriminator loss as well as minimize the source prediction loss and the discrepancy discriminator loss jointly. For source data flow, $\widehat{\theta}_{y_1}$ and $\widehat{\theta}_{y_2}$ in Eq. (11) will minimize the source prediction loss. While in Eq. (12), target data flow only influences the discrepancy discriminator loss. $\widehat{\theta}_{y_1}$ and $\widehat{\theta}_{y_2}$ attempt to maximize the difference in terms of target predictions. It is easy to notice that Eq. (11) and Eq. (12) are not contrary to each other, since they are optimized with respect to different domain data flows. In such way, Eq. (8) can be solved easily by adding a simple GRL as shown in Figure 2. Once the model is well-trained, the feature extractor can derive domain-invariant features with discriminative information preserved, encouraging fine-grained alignment across domains.

4 EXPERIMENT

In this section, we perform an extensive empirical evaluation of the proposed approach with several state-of-the-art (SOTA) deep DA methods on inter twinning moons 2D problems, digits and object cross-domain classification benchmarks.

4.1 Setup

VisDA-2017 [35] is a challenging synthetic to real dataset, which represents a large-scale cross-domain object classification benchmark. It contains over 280K images across 12 categories in the training, validation and test sets. As [39], we choose the training set as source domain which contains 152,397 synthetic images, renderings 3D CAD models from different angles and under different lighting conditions. As for target domain, we choose the validation set collected from MSCOCO [23] that contains 55,388 real images.

ImageCLEF[1] is a dataset for ImageCLEF 2014 domain adaptation challenge, which is organized by selecting 12 common classes from datasets: Caltech-256 (C), ImageNet ILSVRC2012 (I), and PASCAL VOC2012 (P). We use all datasets as three domains and perform six cross-domain tasks.

[1]http://imageclef.org/2014/adaptation

Office-31 [38] is a popular benchmark for visual DA, comprising 31 classes of 4,110 images drawn from three distinct domains: Amazon (A), DSLR (D) and Webcam (W). Specifically, Amazon consists of images downloaded from the online web merchants. DSLR includes high-resolution images captured by a digital SLR camera while Webcam contains low-resolution images recorded by a web camera. We evaluate our method on three cross-domain tasks: A → W, D → W, and W → D as [10].

Digits Datasets contain five standard digital datasets: MNIST [19], MNIST_M [10], USPS [16], Street View House Numbers (SVHN) [31], and synthetic digits dataset (SYN) [10]. They all consist 10 classes of digits. We assess five adaptation scenarios: SVHN → MNIST, SYN → MNIST, MNIST → USPS, USPS → MNIST, and MNIST → MNIST_M.

4.2 Implementation Details

For a fair comparison, we adopt ResNet-101 [13] model pre-trained on ImageNet [37] as base network following the protocol in [39] for VisDA-2017. We regard the pre-trained model as the feature network and substitute the last fully-connected layer by three-layered fully-connected networks as discriminator networks with random initialization. We utilize mini-batch SGD optimizer with momentum 0.9 and learning rate 5×10^{-4}. All layers are updated with the same learning rate as suggested in [39]. In addition, we fix $\lambda = 1$ throughout all experiments in this paper since JADA performs stably under different parameter settings.

For the experiments on ImageCLEF and Office-31, we follow the standard protocols as [9, 34] and evaluate the performance of ResNet-50 model that pre-trained on ImageNet. Similarly, the last fully-connected layer is replaced by three-layered fully-connected networks. The discriminator networks are trained from scratch with learning rate 10 times that of the base learning rate as [34]. And we also adopt mini-batch SGD with momentum 0.9, learning rate 1×10^{-3} and learning rate annealing strategy as [10]. For the experiments on digital benchmarks, we employ the same network architectures for the discriminator and feature generator networks provided by [39]. Also, we adopt mini-batch SGD with momentum 0.9, learning rate 1×10^{-3}.

Note that all the above methods are implemented via Pytorch. For reducing parameter sensitivity and easing the selection of models like [9, 34], we adopt a progressive strategy for the discriminators, gradually increasing λ from 0 to 1 by a schedule [9]: $\lambda_p = \frac{2}{1+exp(-\delta \times p)} - 1$, where $\delta = 10$ is fixed and p is the training process linearly changing from 0 to 1. We adopt all labeled source data and all unlabeled target data and report the average classification accuracy of each task based on 3 random experiments.

4.3 Results

We compare our proposed model against multiple SOTA unsupervised deep DA methods, including DDC [46], DAN [27], JAN [29], RTN [28], JDDA [5], CMD [52], DANN [10], MADA [34], ADDA [45], CoGAN [26], UNIT [25], CyCADA [14], PixelDA [3], HADN [40] and MCD [39]. Note that the presented results of baselines are directly reported from their corresponding papers if the experiment settings are the same.

Table 1: Accuracy(%) on VisDA-2017 for unsupervised DA (ResNet-101).

Method	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg
Source Only	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DAN [27]	87.1	63.0	76.5	42.0	**90.3**	42.9	**85.9**	53.1	49.7	36.3	**85.8**	20.7	61.1
DANN [10]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
MCD [39]	87.0	60.9	**83.7**	64.0	88.9	79.6	84.7	**76.9**	**88.6**	40.3	83.0	25.8	71.9
JADA	**91.9**	**78.0**	81.5	**68.7**	90.2	**84.1**	84.0	73.6	88.2	**67.2**	79.0	**38.0**	**77.0**

Experiment results on VisDA-2017: Table 1 reports the results on VisDA-2017. The first line shows accuracies when the unadapted source classifier is directly applied to target domain. We can clearly observe that our model performs better than the Source Only model in all object categories, while the domain-wise alignment based methods perform worse in several categories, i.e., car and truck. In addition, our model also outperforms the best baseline MCD [39] by a large margin. An interpretation is that there exist large domain-wise and class-wise distribution discrepancies between this challenging synthetic to real image adaptation task. Although task-specific decision boundaries are utilized in MCD, it is still not enough to ideally align source and target due to each domain's characteristics. Hence, it is crucial to jointly conduct domain-wise and class-wise adaptation to align such distributions.

Table 2: Accuracy(%) on ImageCLEF for unsupervised DA (ResNet-50).

Method	I→P	P→I	I→C	C→I	C→P	P→C	avg.
Source Only	74.8	83.9	91.5	78.0	65.5	91.2	80.7
DAN [27]	74.5	82.2	92.8	86.3	69.2	89.8	82.5
RTN [28]	75.6	86.8	95.3	86.9	72.7	92.2	84.9
JAN [29]	76.8	88.0	94.7	89.5	74.2	91.7	85.8
DANN [10]	75.0	86.0	**96.2**	87.0	74.3	91.5	85.0
MADA [34]	75.0	87.9	96.0	88.8	75.2	92.2	85.8
MCD [39]	77.3	89.2	92.7	88.2	71.0	92.3	85.1
JADA	**78.2**	**90.1**	95.9	**90.8**	**76.8**	**94.1**	**87.7**

Table 3: Accuracy(%) on Office-31 for unsupervised DA (ResNet-50).

Method	A → W	D → W	W → D
Source Only	68.4±0.2	96.7±0.1	99.3±0.1
DAN [27]	80.5±0.4	97.1±0.2	99.6±0.1
RTN [28]	84.5±0.2	96.8±0.1	99.4±0.1
JAN [29]	85.4±0.3	97.4±0.2	99.8±0.2
ADDA [45]	86.2±0.5	96.2±0.4	98.4±0.3
DANN [10]	82.0±0.4	96.9±0.2	99.1±0.1
MADA [34]	90.0±0.1	97.4±0.1	99.6±0.1
JDDA [5]	82.6±0.4	95.2±0.2	99.7±0.0
MCD [39]	85.7±0.5	94.7±0.2	99.3±0.2
JADA	**90.5±0.1**	**97.5±0.3**	**100.0±0.0**

Experiment results on ImageCLEF and Office-31: Table 2 and 3 report the accuracies for ImageCLEF and Office-31, respectively. Our model outperforms the comparisons over most tasks. It is desirable that JADA improves the classification accuracy on hard tasks, i.e., A → W where the source and target domains are tremendously different [38]. Note that JADA also outperforms MADA [34], which has multiple discriminators based on the number of classes, whereas JADA only consists of two discriminators. As reported in Table 2, the encouraging results emphasize the importance of joint alignment for deep DA problems, and reveal that JADA is able to learn more transferable features for effective domain adaptation.

Table 4: Accuracy(%) on Digits for unsupervised DA.

Method	SVHN ↓ MNIST	SYN ↓ MNIST	MNIST ↓ USPS	USPS ↓ MNIST	MNIST ↓ MNIST_M
Source Only	67.1	89.7±0.2	79.4	63.4	62.8±0.2
DDC [46]	71.9±0.4	89.9±0.2	-	75.8±0.3	78.4±0.1
DAN [27]	79.5±0.3	75.2±0.1	-	89.8±0.2	79.6±0.2
CMD [52]	86.5±0.3	96.1±0.2	-	86.3±0.4	85.5±0.2
DANN [10]	71.1	90.2±0.2	85.1	73.0±0.2	76.7
JDDA [5]	94.2±0.1	97.7±0.0	-	96.7±0.1	88.4±0.2
ADDA [45]	76.0±1.8	96.3±0.4	-	90.1±0.8	-
CoGAN [26]	-	-	-	89.1±0.8	-
HADN [40]	84.9±0.1	-	91.9±0.1	96.0±0.1	-
PixelDA [3]	-	-	95.9	-	**98.2**
UNIT [25]	90.5	-	96.0	93.6	-
CyCADA [14]	90.4±0.4	-	95.6±0.2	96.5±0.1	-
MCD [39]	96.2±0.4	-	96.5±0.3	94.1±0.3	-
JADA	**96.4±0.2**	**98.6±0.2**	**97.6±0.2**	**97.1±0.3**	92.9±0.2

Experiment results on Digits: Table 4 reports the accuracies for digital datasets. We can see that JADA is superior to competing approaches in most scenarios. It is interesting to observe that the adversarial-based methods (DANN, PixelDA, CyCADA, and MCD) achieve better performance than discrepancy-based methods (DDC, DAN, CMD, JDDA), which proves the importance of utilizing adversarial training process to guide the process of domain adaptation and thus improve the generalization performance. Furthermore, the task-specific decision discrepancy-aware methods such as MCD and the proposed JADA are the current leading approaches for the digital tasks. This demonstrates the exploration of discriminative structures within the data will further lead to a more precise classification performance.

4.4 Empirical Analysis

Inter Twinning Moons 2D Problems: In this experiment, we compare the decision boundaries of Source Only, DANN [10], MCD [39] and JADA on inter twinning moons 2D problems [33]. As shown in Figure 3, red (the upper moon) and green (the lower moon) points represent source class 0 and 1, each of which contains 100 samples. Unlabeled target samples are generated by rotating the distribution of source with an angle of 30°. The dark line is the decision boundary derived by each method. To guarantee the fairness of the experiment, we employ the same network architecture

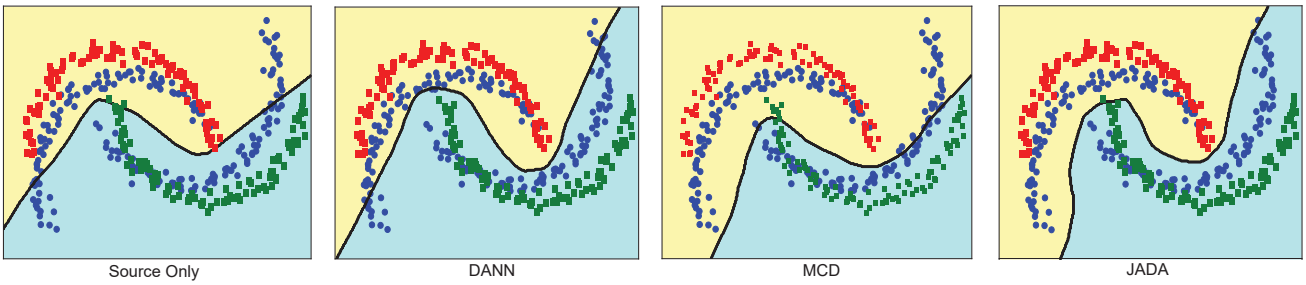

Figure 3: (Best viewed in color.) Comparison of Source Only, DANN, MCD and JADA on inter twinning moons 2D problems. Red and green points denote the source samples of class 0 and 1. Blue points are target samples generated from source samples by rotating 30°. The yellow and light blue regions are classified as class 0 and 1 by the final decision boundary, respectively.

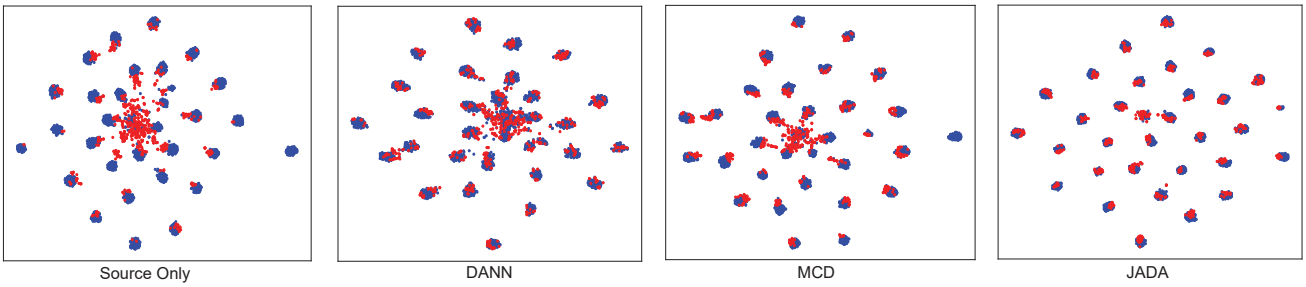

Figure 4: t-SNE [30] visualization of features extracted from task A → W of Office-31 by Source Only, DANN, MCD and JADA. Blue and red points represent the source and target samples, respectively.

in [39], which consists of three-layered fully-connected networks for the feature generator and discriminators.

We observe that Source Only model without adaptation can correctly classify all the source samples, but suffers a significant decrease in the performance on target data. Compared to Source Only model, DANN performs better, since DANN utilizes domain adversarial learning to align source and target domains. However, DANN does not generalize well to the target samples that are far away from the source domain. An interpretation is that only conducting domain-wise alignment may destroy the discriminative structures of the learned features. MCD is able to classify target class 0 accurately, but performs poorly on target class 1. The reasons may rely on the inaccurate prediction of two task-specific classifiers in the beginning. By contrast, due to joint domain-wise and class-wise alignment, JADA could adapt to the target domain nicely and draw a correct decision boundary in almost all regions. The performance of JADA in this experiment proves the superiority to other compared approaches.

Feature Visualization: Figure 4 illustrates the t-SNE [30] embedding of feature representations learned by Source Only, DANN, MCD and JADA for task A → W of Office-31. The source and target samples are not matched well with Source Only features, and better matched with DANN features, since DANN tries to align global distributions across domains. But the category discriminativeness is not preserved well. Compared with DANN, MCD features can be discriminated better. The reason is MCD aims to align source and target class-wisely utilizing the task-specific decision boundaries. However, there exists one cluster of source samples not having the corresponding target samples in the right part of the figure. This phenomenon implies that when target samples are misclassified by the two classifiers simultaneously in the beginning, MCD may not be able to predict these data correctly after convergence. While JADA could align two domains nicely, which indicates the benefit of conducting domain-wise and class-wise alignments jointly.

Misclassified Samples Analysis: To deeply explore the advantages of JADA over other baselines, Figure 5 shows randomly selected misclassified samples of DANN and MCD for the task A → W with respect to the classes "mobile phone" and "ruler", since DANN and MCD perform poorly on these two categories. Especially, MCD misclassifies all the target samples of class "mobile phone", which is consistent with the observation in t-SNE visualization. By observing the training process of MCD, we find all the target samples of class "mobile phone" are predicted wrongly by two task-specific classifiers simultaneously from beginning. This indicates that MCD heavily relies on the correctness of source classifiers. For DANN, it will misclassify the target samples that are much similar to other classes in source domain, which means without considering the discriminative structures, DANN will mix up source and target samples. Therefore, the domain-wise alignment and class-wise alignment are complementary to each other. By capturing the global domain-wise knowledge and preserving the discriminative information, JADA will achieve higher cross-domain prediction accuracies.

Confusion matrices: We draw the confusion matrices in Figure 6 to intuitively illustrate the efficacy of our approach. For the Source Only model, there are many wrong digit predictions. For instance, most samples of class "8" are mistakenly predicted into "3"

Figure 5: Misclassified samples analysis of DANN and MCD for task A → W of Office-31 with respect to classes "mobile phone" and "ruler". Red and black represent the misclassified and correct samples.

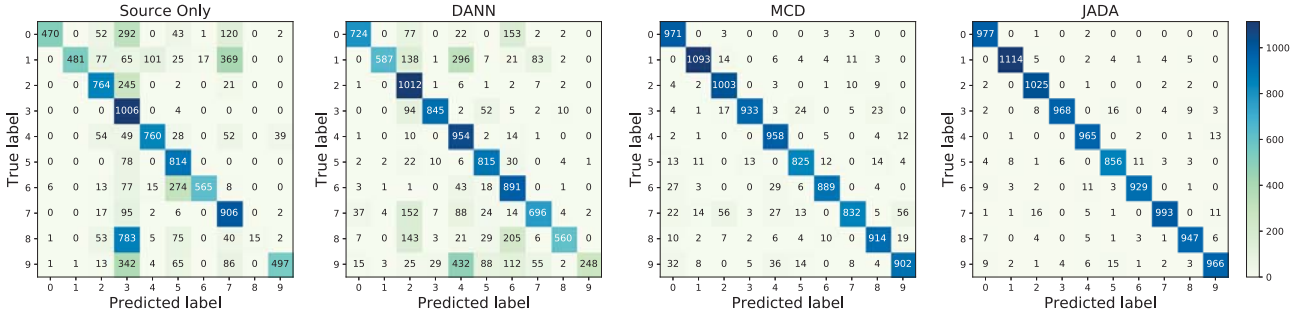

Figure 6: (Best viewed in color.) Confusion matrices for visualization of the performance of Source Only, DANN, MCD and JADA for task USPS → MNIST.

which reveals the tremendously large difference among domains. DANN [10] and MCD [39] perform better, but in some cases it is quite possible to be misclassified, particularly when testing similar digits like "7" and "2", "9" and "4". By contrast, much more right predictions appear in the diagonal using JADA, which proves both the domain-wise and class-wise discrepancies could be effectively mitigated by the proposed method.

Convergence Performance: The optimization of two minimax games in JADA can be achieved simultaneously by optimizing Eq. (8). This enables training JADA efficiently in an end-to-end way by adding a gradient reversal layer as [10, 34]. To verify the convergence performance of JADA, Figure 7 shows the test errors of different methods for task A → W of Office-31. JADA converges faster and performs more stable than one-level minimax game based methods.

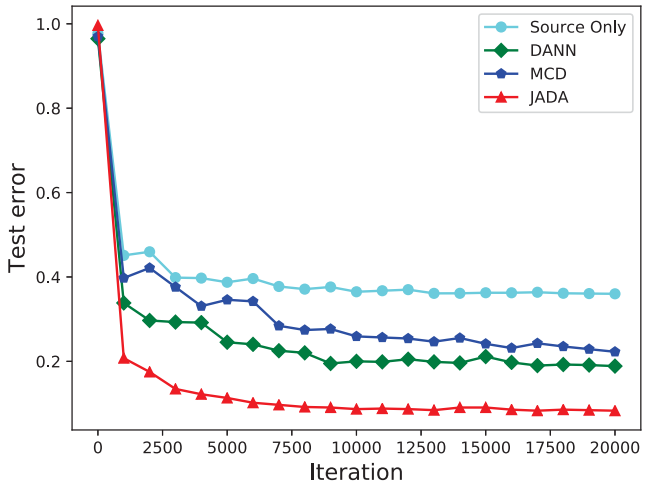

Figure 7: Convergence performance for task A → W.

5 CONCLUSION

In this paper, we propose a novel joint adversarial domain adaptation (JADA) approach for DA by simultaneously capturing the global domain knowledge and exploiting category discriminative structures. Unlike previous adversarial learning based DA methods, JADA jointly minimizes the domain-wise and class-wise distribution discrepancies via an unified adversarial learning process and is more robust to large domain shift. JADA can be optimized in an efficient end-to-end training manner via back-propagation. Empirical evidence demonstrates that JADA has superiority over state-of-the-art methods on several standard cross-domain datasets.

REFERENCES

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79, 1-2 (2010), 151–175.

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *Adv. Neural Inf. Process. Syst. (NIPS)*. 137–144.

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *IEEE Conf. on Comput. Vis. and Pattern Recognition (CVPR)*. 3722–3731.

[4] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. 2018. Partial Transfer Learning with Selective Adversarial Networks. (2018), 2724–2732.

[5] Chao Chen, Zhihong Chen, Boyuan Jiang, and Xinyu Jin. 2018. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. *arXiv preprint arXiv:1808.09347* (2018).

[6] Zhengming Ding, Sheng Li, Ming Shao, and Yun Fu. 2018. Graph adaptive knowledge transfer for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 37–52.

[7] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. [n. d.]. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition.. In *Int. Conf. Mach. Learn. (ICML)*.

[8] Thomas Forgione, Axel Carlier, Géraldine Morin, Wei Tsang Ooi, Vincent Charvillat, and Praveen Kumar Yadav. 2018. An Implementation of a DASH Client for Browsing Networked Virtual Environment. In *26th ACM Int. Conf. on Multimedia (MM '18)*. ACM, 1263–1264.

[9] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Int. Conf. Mach. Learn. (ICML)*. 1180–1189.

[10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res. (JMLR)* 17, 1 (2016), 2096–2030.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Adv. Neural Inf. Process. Syst. (NIPS)*. 2672–2680.

[12] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. 2007. A kernel method for the two-sample-problem. In *Adv. Neural Inf. Process. Syst. (NIPS)*. 513–520.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conf. on Comput. Vis. and Pattern Recognition (CVPR)*. 770–778.

[14] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *Int. Conf. Mach. Learn. (ICML)*. 1994–2003.

[15] Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola. 2007. Correcting sample selection bias by unlabeled data. In *Adv. Neural Inf. Process. Syst. (NIPS)*. 601–608.

[16] Jonathan J. Hull. 1994. A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* 16, 5 (1994), 550–554.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Adv. Neural Inf. Process. Syst. (NIPS)*. 1097–1105.

[18] Abhishek Kumar, Prasanna Sattigeri, Kahini Wadhawan, Leonid Karlinsky, Rogerio Schmidt Feris, Bill Freeman, and Gregory W Wornell. 2018. Co-regularized Alignment for Unsupervised Domain Adaptation. *Adv. Neural Inf. Process. Syst. (NIPS)* (2018), 9345–9356.

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[20] Shuang Li, Shiji Song, and Gao Huang. 2017. Prediction reweighting for domain adaptation. *IEEE Trans. Neural Netw. Learn. Sys. (TNNLS)* 28, 7 (2017), 1682–1695.

[21] Shuang Li, Shiji Song, Gao Huang, Zhengming Ding, and Cheng Wu. 2018. Domain Invariant and Class Discriminative Feature Learning for Visual Domain Adaptation. *IEEE Trans. Image Process. (TIP)* 27, 9 (2018), 4260–4273.

[22] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Interpretable Multimodal Retrieval for Fashion Products. In *26th ACM Int. Conf. on Multimedia (MM '18)*. ACM, 1571–1579.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis. (ECCV)*. Springer, 740–755.

[24] Zehang Lin, Zhenguo Yang, Runwei Situ, Feitao Huang, Jianming Lv, Qing Li, and Wenyin Liu. 2018. Improving maximum classifier discrepancy by considering joint distribution for domain adaptation. In *International Conference on Web Information Systems Engineering*. Springer, 253–268.

[25] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Adv. Neural Inf. Process. Syst. (NIPS)*. 700–708.

[26] Ming-Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In *Adv. Neural Inf. Process. Syst. (NIPS)*. 469–477.

[27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. 2015. Learning transferable features with deep adaptation networks. In *Int. Conf. Mach. Learn. (ICML)*. 97–105.

[28] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2016. Unsupervised domain adaptation with residual transfer networks. In *Adv. Neural Inf. Process. Syst. (NIPS)*. 136–144.

[29] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2017. Deep Transfer Learning with Joint Adaptation Networks. In *Int. Conf. Mach. Learn. (ICML)*. 2208–2217.

[30] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res. (JMLR)* 9, Nov (2008), 2579–2605.

[31] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. (2011).

[32] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. 2011. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw. (TNN)* 22, 2 (2011), 199–210.

[33] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron J Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res. (JMLR)* 12 (2011), 2825–2830.

[34] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-adversarial domain adaptation. In *AAAI Conf. Art. Intell. (AAAI)*.

[35] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. 2017. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924* (2017).

[36] Fan Qi, Xiaoshan Yang, and Changsheng Xu. 2018. A Unified Framework for Multimodal Domain Adaptation. In *26th ACM Int. Conf. on Multimedia (MM '18)*. ACM, 429–437.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. 2015. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis. (IJCV)* 115, 3 (2015), 211–252.

[38] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *Eur. Conf. Comput. Vis. (ECCV)*. 213–226.

[39] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE Conf. on Comput. Vis. and Pattern Recognition (CVPR)*. 3723–3732.

[40] Rui Shao, Xiangyuan Lan, and Pong C. Yuen. 2018. Feature Constrained by Pixel: Hierarchical Adversarial Deep Domain Adaptation. In *26th ACM Int. Conf. on Multimedia (MM '18)*. ACM, 220–228.

[41] Rui Shu, Hung Hai Bui, Hirokazu Narui, and Stefano Ermon. 2018. A DIRT-T Approach to Unsupervised Domain Adaptation. *international conference on learning representations (ICLR)* (2018).

[42] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *AAAI Conf. Art. Intell. (AAAI)*. 2058–2065.

[43] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Eur. Conf. Comput. Vis. (ECCV)*. 443–450.

[44] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *IEEE Conf. on Comput. Vis. and Pattern Recognition (CVPR)*. IEEE.

[45] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *IEEE Conf. on Comput. Vis. and Pattern Recognition (CVPR)*, Vol. 1. 4.

[46] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv:1412.3474* (2014).

[47] Vedran Vukotić, Christian Raymond, and Guillaume Gravier. 2016. Multimodal and Crossmodal Representation Learning from Textual and Visual Features with Bidirectional Deep Neural Networks for Video Hyperlinking. In *2016 ACM Workshop on Vision and Language Integration Meets Multimedia Fusion (iV&L-MM '16)*. ACM, 37–44.

[48] Cheng Wang, Mathias Niepert, and Hui Li. 2019. RecSys-DAN: Discriminative Adversarial Networks for Cross-Domain Recommender Systems. *IEEE Transactions on Neural Networks* (2019), 1–10.

[49] Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing* 312 (2018), 135–153.

[50] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Online Asymmetric Metric Learning with Multi-Layer Similarity Aggregation for Cross-Modal Retrieval. *IEEE Trans. Image Process. (TIP)* (2019).

[51] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Adv. Neural Inf. Process. Syst. (NIPS)*. 3320–3328.

[52] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811* (2017).

[53] Junbao Zhuo, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2017. Deep Unsupervised Convolutional Domain Adaptation. In *25th ACM Int. Conf. on Multimedia (MM '17)*. ACM, 261–269.