Class Consistency Driven Unsupervised Deep Adversarial Domain Adaptation

Sayan Rakshit, Ushasi Chaudhuri, Biplab Banerjee, Subhasis Chaudhuri
IIT Bombay, India
{sayan1by2, ushasi2cool, getbiplab} @gmail.com, sc@ee.iitb.ac.in

Abstract

In unsupervised deep domain adaptation (DA), the use of adversarial domain classifiers is popular in learning a shared feature space which reduces the distributions gap for a pair of source (with training data) and target (with only test data) domains. In the new space, a classifier trained on source training data is expected to generalize well for the target domain samples.

We hypothesize that such a feature space obtained by aligning the domains globally ignores the category level feature distributions. This, in turn, leads to erroneous mapping for fine-grained classes. Besides, the discriminativeness of the shared space is not explicitly addressed. In order to resolve both the issues, we propose a novel adversarial approach which judiciously refines the space learned by the domain classifier by incorporating class level information. We follow an ensemble classifiers based approach to model the source domain and introduce a novel consistency constrain on the classifier's outcomes when evaluated on a held-out set of target domain samples. We further leverage the ensemble learning strategy during the inference, as opposed to the existing single classifier based methods. We find that our deep DA model is capable of producing a compact and better domain aligned feature space. Experimental results obtained on the Office-Home, Office-CalTech, MNIST-USPS, and a remote sensing dataset confirm the superiority of the proposed approach.

1. Introduction

Traditional classifier models inherently assume that the probability distributions governing the training and test samples are consistent in nature, following the probably approximately correct (PAC) assumptions of the statistical learning theory [38]. However, such a restriction is difficult to satisfy in practise considering the inherent ambiguities in capturing, in our case visual data, from diverse sources. Besides, it is non-trivial to annotate training samples manually for all the sources separately since the capability of manual labeling is unlikely to match the rapid data generation rate. This prompts for the exploration of the notion of transferring supervised inference models trained on one source to others with lack of annotations. To this end, the paradigm of unsupervised DA [41, 34, 8, 35] is highly explored which is defined as follows: Given a *source* domain S equipped with ample amount of labeled training data, the task is to learn a classifier which will also perform well for test samples obtained from a related but different target domain T.

Amongst different approaches towards solving DA, subspace learning techniques [24] are of particular interest since they directly learn a shared feature space for both the domains where the marginal distributions of the cross-domain data can be related to. In this regard, several ad-hoc feature transformation based strategies are present in the literature which have shown impressive performance with hand-crafted features [26]. Subsequently, there have been numerous endeavors with the deep learning models which perform the task of domain invariant feature learning directly from the data within an end-to-end framework [32]. While some of the approaches in this regard propose to reduce the domain difference in terms of the minimization of some higher-order statistics between S and T, the notion of domain classification [8] is also a popular paradigm. It aims at learning a common subspace where the classifier confusion is constrained to be maximized through an adversarial training process. More evidently, assuming that the labels of the samples from S and T be denoted by 0 and 1, respectively, the domain classifier is trained to maximize a typical cross-entropy type loss in a multi-layer neural network setup. This, in essence, enforces the learned features through a *generator/feature extractor* sub-network to be independent of the domains. Focusing on the ultimate goal of domain independent data classification, several techniques concurrently model a source specific classifier in this subspace. Since the domain classifier aligns the domains, it is assumed that the trained source classifier is capable of classifying the target samples highly accurately [18, 35].

However, this idea has two potential drawbacks: i) it does not explicitly guarantee discriminativeness in the obtained features: although a source specific classifier is trained to look after the between class separation, the notion of intra-class compactness is largely overlooked, and ii) the correlation between the domains at a finer level is not carefully explored: the domain classifier does not explicitly incorporate the class level information while globally aligning S and T. As a consequence, such an approach largely fails to propagate the complicated dependencies among the classes from the original to the learned feature space.

Inspired by the aforementioned discussions, we propose the following extensions to the said adversarial deep domain classification based unsupervised DA model: i) instead of modeling one source classifier with all the training samples, we consider to deploy a bagging based ensemble classifiers trained on mutually-exclusive training subsets. This committee of classifiers is undoubtedly a better choice to model the visual classes considering their complex distributions and a decision fusion in this regard is expected to assure better generalization. Note that our inference in T is also guided by the ensemble voting strategy. ii) At the same time, we deploy a subset of target domain samples for the class level domain alignment purpose. Particularly, we consider the responses of all the source classifiers on these target data and specifically demand these response vectors to be highly similar during the course of training. We term the respective loss as the *logit homogeneity loss* for target samples. This desiderata has two-fold advantages: i) in order to ensure consistency in the classifier committee's outputs on the held-out target set, the feature extractor sub-network implicitly enforces the samples to be class-wise concentrated, and ii) while the domain classifier aligns S and T globally, the proposed simultaneous class scale alignment of the target samples entails the notion of semantic consistency between the domains. As a whole, we follow an alternate optimization strategy to ensure that the feature space learned by our improved domain classifier to be more robust and compact where the global domain alignment is subsequently fine-tuned locally using class distributions. We summarize the noteworthy contributions of the work in the following:

i) We propose a novel adversarial loss driven unsupervised deep DA framework which efficiently exploits the class level information along with the domain classification loss to better associate both the domains. In this respect, we wisely deploy the notion of ensemble learning within the DA framework. ii) We theoretically validate our model with respect to established generalization bounds. iii) Experimentally, we find that our model performs better or commensurately with recent DA approaches on the challenging Office-Home, Office-CalTech, MNIST-USPS, and a remote sensing dataset.

2. Related works

Broadly, the existing techniques for DA rely on reweighting the source training samples to distinctively reflect their counterparts in the target distributions [27] or obtaining a transformation in a much lower-dimensional manifold that makes the target features indistinguishable from the source. Notice that the instance re-weighting based approaches are considered to be a rather restricted form of domain-shift or sample selection bias and cannot be generalized to critical scenarios. On the other hand, the Geodesic distance based methods [10, 11] consider the domains to be points on a typical Grassmannian manifold and bridge the domain gap by projecting S and T on the points along the geodesic path or find a closed form linear map which projects samples from S onto the distributions of T. Likewise, CORAL [31] minimizes the domain-shift in terms of the Euclidean distance between the second-order statistics governing S and T whereas Log-CORAL [42] considers their Log-Euclidean distance on a Riemannian manifold. A detailed survey of DA for visual recognition can be obtained in [41, 26].

The deep neural network based approaches, on the other hand, learn domain invariant features using supervised models, deep auto-encoders [5], and recently generative adversarial networks (GAN) [18, 35]. As aforementioned, these methods are designed to minimize a classification loss while optimizing a measure for ensuring domain consistency. An extension of CORAL is deep CORAL [33] which integrates a convolutional network (CNN) based feature extractor in the CORAL framework for end-to-end learning. As opposed to this type of loss, several methods have deployed the maximum mean discrepancy (MMD) for comparing distributions in Hilbert space: DDC [36], DAN [20].

Other competing methods are in favor of adversarial loss functions to reduce the domain difference: learning a representation space that is simultaneously discriminative for source labels while being insensitive to individual domain properties. The gradient reversal algorithm (RevGrad) [8] is of particular interest in this regard: it treats the paradigm of domain invariant feature learning as a binary classification problem and directly maximizes the classification loss by reversing the gradient. ADDA [35] learns discriminative representations by exploring the labels of the source domain and training an inverted label-GAN for domain confusion. Besides, some of the recent techniques rigorously incorporate the generative models (mostly GANs) in the feature learning process: coupled-GAN [19], duplex-GAN [15], adversarial feature augmentation [40]. Cycle-GAN [43] endorses a cycle consistency loss to model the mapping $S \rightarrow T \rightarrow S$ using identical functions. Theoretically speaking, given the three factors provided by [1] for bounding the adaptation loss in a typical DA framework, all these adversarial adaptation methods focus on reducing the domain divergence as reducing the joint generalization error for both S and T is intrinsically hard.

Instead of dealing with the binary domain classifier, a few very recent techniques [29, 23] deploy the idea of maximizing the disagreement between a pair of source domain classifiers on target samples to guide the feature extractor in reducing the domain gap. Ideally, the classifier pair's disagreement highlights potentially confusing target domain samples and the feature extractor learns *better* representations for them. However, such techniques may fail for partially overlapping target classes as both the source classifier's outputs are likely to signify mis-classifications and the models fail to identify and correct the same. As a consequence, the feature space may not turn out to be highly overlapping between the domains as desired. As opposed to them, we aim at alternately maximizing the domain classifier's error (for obtaining the domain invariant space by global domain alignment) and minimizing the average pairwise difference between the classifier committee's responses on held-out target set (for domain realignment according to the class distributions in the domain invariant space) in an end-to-end model. Hence, we ensure a domain independent feature space first and subsequently revamp the space to be semantically coherent. Besides, we distill the advantages of ensemble learning effectively in our model.

3. Method

3.1. Preliminaries

Let $\mathbb{X}^S = \{\mathbf{x_i^S}, y_i^S\}_{i=1}^{N^S}$ on $X^S \otimes Y^S$ denote the training samples from S where $\mathbf{x_i^S} \in X^S$ can be images or descriptors extracted from the images each denoting one of the $y_i^S \in Y^S = \{1, 2, \dots, M\}$ visual categories. In contrast, $\mathbb{X}^T = \{\mathbf{x_j^T}\}_{j=1}^{N^T}$ on X^T represents the set of unlabeled test samples from T. We assume that both S and T can be modeled in terms of unknown underlying distributions $\mathcal{S}$ and $\mathcal{T}$, respectively, with $\mathcal{S}(X^S) \neq \mathcal{T}(X^T)$.

Under this setup, the subspace learning based unsupervised DA paradigm seeks to model a mapping $f_C : X^S \rightarrow Y^S$ in a new feature space such that f_C performs well in predicting labels for samples from $\mathcal{T}$. In order to realize f_C, we propose a model with three components: i) a feature extractor f_G with parameters θ_G: $\tilde{X}^{S/T} = f_G(X^{S/T}; \theta_G)$, ii) a committee of N source domain classifiers $\{f_{C_n}^S\}_{n=1}^N$ with parameters θ_{C_n}: $Y^S = f_{C_n}^S(\tilde{X}^S; \theta_{C_n})$, iii) a binary domain classifier f_D with parameter set θ_D. In this respect, let $y^D \in Y^D = \{0, 1\}$ denote the domain labels for samples coming from $\mathcal{S}$ and $\mathcal{T}$, respectively. Given that, $X^D = [X_{HO}^S, X_{HO}^T]$ represents a mixture of samples from the domains with $X_{HO}^S \subseteq X^S$ and $X_{HO}^T \subseteq X^T$. The underlying mapping function can be interpreted as $Y^D = f_D(\tilde{X}^D; \theta_D)$.

On the other hand, we assume that $f_{C_n}^S$s are trained on N non-overlapping subsets $\{\mathbb{X}_n^S\}_{n=1}^N$ of $\mathbb{X}^S$. Finally for a given $\mathbf{x_k^T} \in X_{HO}^T$, we denote the M- dimensional logit vector corresponding to $f_{C_n}^S$ by $\mathbf{l_n^k}$. In the following, we discuss the training and inference stages of the proposed model in detail. A depiction of the proposed framework can be found in Fig.1.

3.2. Training & inference

As already mentioned briefly, we optimize a novel multi-task loss function for obtaining $[\theta_G, \{\theta_{C_n}\}_{n=1}^N, \theta_D]$. In particular, we are interested in i) a cross-entropy type domain classification loss for f_D which is to be maximized given $\tilde{X}^D$, ii) N separate cross-entropy losses for $f_{C_n}^S$s which are deemed to be minimized given the learned representations of $\mathbb{X}_n^S$s, iii) mean pairwise difference between the logits obtained from $f_{C_n}^S$s for $\tilde{X}_{HO}^T$ which is also to be minimized. iv) An additional norm induced constraint on $\tilde{X}^D$ which is simultaneously minimized in order to control any unwanted feature space diversion.

Domain classifier loss: The prime motivation behind maximizing the cross-entropy loss for f_D is to obtain the coarse level domain invariant feature space through the modeling of f_G which, to a certain extent, satisfies the requirement of $\mathcal{S} \approx \mathcal{T}$. In order to obtain that, the corresponding loss measure, given (X^D, Y^D), can be mentioned as:

$$\mathcal{L}_D(\theta_D, \theta_G) = -\mathbb{E}_{\tilde{\mathbf{x}}_\mathbf{m} \in \tilde{X}_{HO}^S} \log(f_D(\tilde{\mathbf{x}}_\mathbf{m})) - \mathbb{E}_{\tilde{\mathbf{x}}_\mathbf{m} \in \tilde{X}_{HO}^T} \log(1 - f_D(\tilde{\mathbf{x}}_\mathbf{m})) \tag{1}$$

At the same time, we penalize any undesirable variation of the projected feature embeddings by explicitly controlling the feature norm. This can act as a regularizer on the obtained features:

$$\mathcal{L}_R(\theta_G) = ||\tilde{X}^D||_F^2 \tag{2}$$

Source classifier loss: $f_{C_n}^S$s are essentially multi-class classifiers which produce the M-dimensional class probability distributions given $\tilde{X}_n^S$s. Optimizing these classifiers on S ensures better between-class separation in the learned feature space. For the n^{th} classifier, we denote the respective softmax cross-entropy loss measure by:

$$\mathcal{L}_n^S(\theta_G, \theta_{C_n}) = -\mathbb{E}_{(\tilde{\mathbf{x}}_\mathbf{m}, \mathbf{y_m}) \in \mathbb{X}_n^S} \sum_{y_m \in \mathbf{y_m}, p_m \in \mathbf{p_m}} y_m \log(p_m) \tag{3}$$

where $\mathbf{p_m}$ denotes the softmax class distributions vector for a given $\tilde{\mathbf{x}}_\mathbf{m}$ and $\mathbf{y_m}$ represents its label in the one-hot format. As mentioned, this is a bagging based ensemble model where the N classifiers specialize on ideally non-overlapping areas of the feature space. Additionally, the classifiers play a major role in better aligning $\mathcal{S}$ and $\mathcal{T}$ (see below).

Logit homogeneity loss for the held-out target set: Let us recapitulate that the domain classifier focuses on orienting the entire data distributions of $\mathcal{S}$ and $\mathcal{T}$ which does not

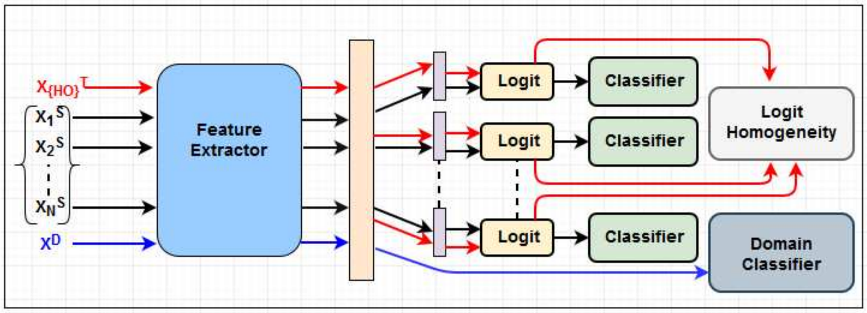


Figure 1: The proposed unsupervised DA framework without back-propagation directions. The variables are mentioned in Section 3.1.

comply to the individual class properties. This is important given the unconstrained multi-modal nature of the visual categories where several classes may overlap in the feature space. To this end, the proposed logit homogeneity loss helps in obtaining a compact domain invariant feature space considering these aspects judiciously.

Given unlabeled X_{HO}^T from $\mathcal{T}$, we propagate the respective learned embeddings $\tilde{X}_{HO}^T$ through all the source-centric classifiers $f_{C_n}^S$s simultaneously and record their logit outcomes. Remember that the logits represent real-valued unnormalized class scores for the samples. We consider logits over the softmax class assignment probabilities here in order to prevent any trivial solution that may occur with typically very small softmax values. With this background discussion, we first define the logit homogeneity loss and further explain how it resolves the aforesaid bottlenecks of the standard approach. Given the logit representations $\mathbf{l_{n1}^k}$ and $\mathbf{l_{n2}^k}$ for a given $\tilde{\mathbf{x}}_\mathbf{k}^\mathbf{T}$ obtained from two different classifiers of the ensemble where $n1, n2 \in \{1, 2, \ldots, N\}$ and $n1 \neq n2$, the concerned loss measure is:

$$\mathcal{L}_T(\theta_G, \{\theta_{C_n}\}_{n=1}^N) = \mathbb{E}_{\tilde{\mathbf{x}}_\mathbf{k}^\mathbf{T} \in \tilde{X}_{HO}^T} \frac{1}{N} \sum_{n1,n2} ||\mathbf{l_{n1}^k} - \mathbf{l_{n2}^k}||_F^2 \tag{4}$$

Now focusing on the main implications of reducing $\mathcal{L}_T$ on the domain invariant space, it can be observed that:

- Ideally, we enforce the class score distributions of $f_{C_n}^S$s on $\mathcal{T}$ to be similar along with reducing their classification errors on $\mathcal{S}$ ($\mathcal{L}_n^S$s) in parallel. It signifies the fact that f_G, in turn, is directed to produce highly concentrated class-wise embeddings for both the domains in the learned feature space to assure the classifier committee's consistency.

- Specifically from the point of view of $\mathcal{T}$, for samples belonging to a given class with high confidence, it is expected that the outcomes of the trained $f_{C_n}^S$s in terms of the class assignment scores should largely coincide. In contrast, this may not be the case for potentially confusing samples. Optimizing $\mathcal{L}_T$ in the learned feature space in this regard basically finds a better mapping for such data items.

Optimization: We follow the standard alternate gradient descent (GD) based optimization strategy to obtain the network parameters. The following two sub-problems are optimized simultaneously until convergence where the λs denote the term weights. Specifically notice that we perform the class-wise domain alignment in the globally domain confused space. Additionally, we enforce to learn the θ_{C_n}s on $\mathcal{S}$ and $\mathcal{T}$ together for better semantic association between domains. We obtain θ_D in Eq.5 practically by applying the RevGrad algorithm [8] which inserts a gradient reversal layer between the shared feature space and domain classifier. Standard version of GD is used in the other cases.

$$\min_{\theta_G} \max_{\theta_D} \lambda_1 \mathcal{L}_R - \lambda_2 \mathcal{L}_D \tag{5}$$

$$\min_{\theta_G, \{\theta_{C_n}\}_{n=1}^N} \sum_{n=1}^{N} \mathcal{L}_n^S + \lambda_3 \mathcal{L}_T \tag{6}$$

Inference: During testing, we follow the traditional decision fusion approach over the outputs of the source classifiers. Specifically, the target samples from $\mathbb{X}^T$ are propagated through the feature extractor as well as the classifiers and the logit scores for the classes are recorded separately for the classifiers. A max-pooling is subsequently performed to fuse the outcomes of the classifiers on which the softmax operation is carried out. Finally, the class with the highest softmax probability is assigned to a given target sample.

3.3. Theoretical insights

In this section, we draw a connection between the proposed method and the DA theory introduced in [1] which relates the error bounds on $\mathcal{S}$ ($R_\mathcal{S}(h)$) and $\mathcal{T}$ ($R_\mathcal{T}(h)$), respectively, for a given hypothesis $h \in \mathcal{H}$. Precisely, there exist two different distance measures to quantify the gap between $\mathcal{S}$ and $\mathcal{T}$: i) $\mathcal{D}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ [1], which can be inferred as the discrepancy between the quality of a classifier on the two domains, and ii) $\mathcal{D}_{\mathcal{H}}(\mathcal{S},\mathcal{T})$ for measuring the domain divergence [2]. Both these distance measures can be compared as follows:

$$\begin{aligned} R_{\mathcal{T}}(h) &\leq R_{\mathcal{S}}(h)+\mathcal{D}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S},\mathcal{T})+\alpha \\ &\leq R_{\mathcal{S}}(h)+\mathcal{D}_{\mathcal{H}}(\mathcal{S},\mathcal{T})+\alpha \end{aligned} \tag{7}$$

where α defines the shared error of the optimum joint hypothesis. Since $\mathcal{D}_{\mathcal{H}}$ upper bounds the $\mathcal{H}\Delta\mathcal{H}$ distance, we find it intuitive to analyze the stability of our method in terms of the $\mathcal{H}$ distance. The actual functional form for $\mathcal{D}_{\mathcal{H}}(\mathcal{S},\mathcal{T})$ can be put forward as:

$$\begin{aligned} \mathcal{D}_{\mathcal{H}}(\mathcal{S},\mathcal{T}) = 2\,\mathbf{sup}_{h\in\mathcal{H}}|[\mathop{\mathbb{E}}_{\mathbf{x^S}\in\mathcal{S}}\mathbb{I}(h(\mathbf{x^S})\neq 1)- \\ \mathop{\mathbb{E}}_{\mathbf{x^T}\in\mathcal{T}}\mathbb{I}(h(\mathbf{x^T})\neq 1)]| \end{aligned} \tag{8}$$

In reality, $\mathcal{D}_{\mathcal{H}}$ highlights empirically the error of the domain classifier, which needs to be maximized in order to obtain the domain invariant feature space. In our case, the proposed framework ensures better domain alignment in terms of a joint coarse to fine level correspondence focusing on both the global and class level agreements. Precisely, $\mathcal{D}_{\mathcal{H}}$ turns out to be:

$$\mathcal{D}_{\mathcal{H}} = \min_{\theta_G}\max_{\theta_D} \lambda_3\mathcal{L}_T - \lambda_2\mathcal{L}_D \tag{9}$$

Nonetheless, it is ensured that a good representation space obtains low values for both the domain specific classifier as well as $\mathcal{D}_{\mathcal{H}}$ simultaneously. Notice that our model uses a committee of classifiers in the source domain and the classifier combination strategy is adopted during inference as well. Since the generalization bound of an ensemble classifiers is always upper bounded by the standalone hypothesis, it means the domain specific error terms are logically minimized in our case thus providing a tighter and better interpretation of $R_{\mathcal{T}}$.

4. Experiments

We detail the quantitative and qualitative evaluations to validate the proposed method in numerous ways in this section.

4.1. Dataset

We carry out experiments in three different scenarios: 1) **Object recognition**: We consider the Office-Home [39] and the Office-CalTech [10] dataset for this purpose. The Office-Home dataset consists of four visual domains, each consisting of images from 65 object categories which results in a total of $15,500$ images. In particular, the domains include, *Art*: (**A**), *Clipart*: (**C**), *Product*: (**P**), and *Real world*: (**Rw**), respectively. We consider all possible combinations: $\mathbf{A} \leftrightarrow \mathbf{C}$, $\mathbf{A} \leftrightarrow \mathbf{P}$, $\mathbf{A} \leftrightarrow \mathbf{Rw}$, $\mathbf{C} \leftrightarrow \mathbf{P}$, $\mathbf{C} \leftrightarrow \mathbf{Rw}$, and $\mathbf{P} \leftrightarrow \mathbf{Rw}$. On the other hand, Office-CalTech is created by selecting 10 shared categories between Office-31 [28] (*Amazon*: (**A**), *Webcam*: (**W**), and *DSLR*: (**D**)) and *CalTech-256* (**C**). It allows 12 possible transfer tasks given the four domains: $\mathbf{A} \leftrightarrow \mathbf{W}, \mathbf{A} \leftrightarrow \mathbf{D}, \mathbf{D} \leftrightarrow \mathbf{W}, \mathbf{C} \leftrightarrow \mathbf{W}, \mathbf{C} \leftrightarrow \mathbf{D}, \mathbf{C} \leftrightarrow \mathbf{A}$. 2) **Digit recognition**: We deal with the MNIST and USPS dataset which contain white digits on a solid black background. We consider two different testing protocols: the first one (P1) consists of sampling 2000 MNIST and 1800 USPS images while the second one (P2) uses the full MNIST and USPS training-test sets as followed by [3], respectively [20, 17]. We test on **MNIST** $\leftrightarrow$ **USPS** for P1 and **MNIST** $\rightarrow$ **USPS** for P2. The USPS images are resized to 28×28 (shape of MNIST images) beforehand to maintain consistency. 3) **Remote sensing image classification**: Finally, we consider the important problem of hyperspectral remote sensing (RS) image pixel classification for a pair of spatially disjoint geographical areas. In particular, we deal with the benchmark Botswana dataset [12] (Fig.2) acquired by the Hyperion sensor of the EO-1 satellite over a 1476×256 pixel study area located at the Okavango Delta, Botswana on May, 2001. Source (2538 pixels) and target (1252 pixels) domain samples are collected from two spatially disjoint sub-areas within the study area from 14 different land-cover classes. 10 out of original 220 bands selected by an ad-hoc feature selection strategy [4] are finally considered to represent the pixels.

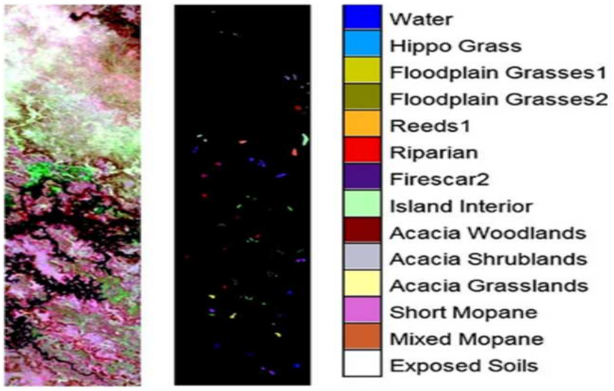


Figure 2: A false color composite (FCC) of the study area, ground-truth sites for S and T from spatially disjoint areas, and the land-cover classes for Botswana dataset.

4.2. Design protocols

Office-Home and Office-CalTech: We rely on the source domain fine-tuned Imagenet pre-trained Resnet-50 [13] features extracted from the last feature layer (*pool5*). All the necessary pre-processing stages are carried out on the images beforehand. Subsequently, f_G consists of two fully-connected (fc) layers coupled with $\mathbf{relu}(\cdot)$ non-linearity and **dropout** units and having dimensions 1000 and 128, respectively. Batch-normalization is used after each fc layer.

MNIST and USPS: A common CNN based feature extractor is used for both these dataset. The CNN structure followed is: (**conv1** → **max-pool1**) → (**conv2** → **max-pool2**) → **fc1** → **fc2** and the final feature dimensions are 128. Batch-normalization and **relu** non-linearity are used after the blocks. The fc layers are also followed by **dropout** units.

RS dataset: The spectral bands are used as the input features. Subsequently, two over-complete fully-connected layers endowed with batch-normalization and **relu** non-linearity are used for constructing f_G with the final feature space having 20 dimensions.

In all cases, the considered classifiers are represented by fc layers with the required number of output units. All the classifiers are trained with the standard cross-entropy loss.

Training protocols: Training is carried out using the Adam optimizer [16] with a learning rate of 0.001 and a batch size of 20. We report the average overall accuracy as the performance measure on the target domain. In order to select the λs, we find that too large or too small λ_2, λ_3 (corresponding to the domain classifier and the target logit-homogeneity loss) lead to poor classification performance and henceforth we set $\lambda_2 = \lambda_3 = 1$. Likewise, we consider different values of λ_1 in the range $[1, 0.0001]$ for all the dataset and report the best performance. X^S_{HO} and X^T_{HO} are constructed using the entire $\mathbb{X}^S$ and $\mathbb{X}^T$. Finally, we report the results considering two source classifiers for the ensemble, but have done some sensitivity analysis, apart from other factors, on the number of classifiers for **MNIST** → **USPS** (P1) (Section 4.4).

4.3. Comparative analysis [1]

Office-Home: Table 1 details the comparative performance analysis of the proposed approach with respect to a number of recent DA techniques. All the methods are evaluated on the same initial Resnet-50 features and identical evaluation protocol is followed. In the trivial case, it is found that when the source classifier is directly applied to the target domain, a mean classification performance of 46.1% can be obtained. Subsequently, the standard adversarial learning based methods are tested which produce the mean average accuracy in the range between: DAN- 56.3% and JAN [22]-58.3%. The CDAN method [21], on the other hand, incorporates discriminative class information in the adversarial training process and outputs a mean classification performance of 63.8%. Our method, which is also based on incorporating the idea of discriminative class information within the adversarial training process using ensemble learning, reports a performance of 64.3%, outperforming the rests.

Office-CalTech: Similar to Office-Home, we compare the performance of our technique on Office-CalTech with four techniques where Resnet-50 features are used (Table 2). In the base scenario, a multi-class support vector machine (SVM) trained on the source domain is applied to the target data without adaptation which produces an accuracy of 91.3%. Subsequently, two ad-hoc techniques are also deployed: subspace alignment (SA) [31] where a mapping function is modeled to project the source domain samples to the target domain and CORAL [7], respectively and we find CORAL outperforms SA slightly by 0.9%. The very recent RWA [37] employs the idea of finding stable labels in the target and pose it as a optimum random walk problem and it produces a mean accuracy of 96.3%. We are better than RWA in eight out of twelve cases and overall we further enhance the average classification performance by 0.5% as compared to RWA, thus obtaining the new state-of-art.

MNIST-USPS: As already mentioned, we consider two experimental scenarios P1 and P2 for MNIST and USPS. Apart from DANN, ADDA and DSN, here we consider six new techniques for comparison purpose all of which are based on adversarial training. It is found that the use of discriminative class information indeed helps the training process for all the experimental cases (Table 3). Moreover, the performance of these methods are found to be highly similar to each other. In contrast, we observe that our approach outperforms the others by at least 0.8% for **MNIST** → **USPS** (P1) and by at least 0.2% for **MNIST** → **USPS** (P2) while providing comparable performance for **USPS** → **MNIST** with ADDA, Generate to adapt and at the same time beating the others.

RS dataset: Table 4 summarizes the performance comparison of our technique with respect to four representative techniques from the literature. Owing to high overlap between a number of land-cover classes in the spectral domain, none of the techniques are able to produce very high accuracy in this regard. Precisely, TCA [25], GFK [10], and SA [31] can produce a maximum of 70% classification rate. We are able to further extend the same by 7.5% to obtain an overall classification performance of 77.5%. Finally, we see that our model outperforms the DANN [9] method by 4.5% in this case.

4.4. Critical discussions

Here we consider the **MNIST** → **USPS** (P1) case to qualitatively assess our model.

Visualization: Fig.3 depict the t-SNE plots for MNIST and USPS before and after feature adaptation. In this respect, we also highlight the effect of $\mathcal{L}_T$. It can be observed that our model obtains better class overlapping (Fig.3(e)) than the standard scenario (domain confusion + source do-

[1] Results are taken from the respective papers

Method	$A \rightarrow C$	$A \rightarrow P$	$A \rightarrow Rw$	$C \rightarrow A$	$C \rightarrow P$	$C \rightarrow Rw$	$P \rightarrow A$	$P \rightarrow C$	$P \rightarrow Rw$	$Rw \rightarrow A$	$Rw \rightarrow C$	$Rw \rightarrow P$	Average
Resnet (source only) [13]	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN [25]	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN [9]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN [22]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN-RM [21]	49.2	64.8	72.9	53.8	62.4	62.9	49.8	48.8	71.5	65.8	56.4	79.2	61.5
CDAN-M [21]	50.6	65.9	73.4	**55.7**	62.7	64.2	51.8	**49.1**	74.5	68.2	**56.9**	**80.7**	62.8
CDAN [21]	49.0	69.3	74.5	54.4	66.0	68.4	**55.6**	48.3	75.9	**68.4**	55.4	80.5	63.8
Ours	**53.2**	**73.1**	**77.2**	55.2	**66.2**	**68.8**	52.3	48.5	**76.8**	67.2	54.4	79.8	**64.3**

Table 1: A comparative study on Office-Home dataset using Resnet-50 features (in %).

Method	$A \rightarrow C$	$A \rightarrow D$	$A \rightarrow W$	$C \rightarrow A$	$C \rightarrow D$	$C \rightarrow W$	$D \rightarrow A$	$D \rightarrow C$	$D \rightarrow W$	$W \rightarrow A$	$W \rightarrow C$	$W \rightarrow D$	Average
Source-SVM	89.4	92.3	89.7	93.6	91.0	87.6	91.2	86.7	97.9	90.5	86	99.9	91.3
SA [31]	88.9	91.8	89.8	93.4	90.3	90.2	91.4	85.8	97.8	90.7	85.4	99.8	91.3
CORAL [7]	89.2	92.2	91.9	94.1	92.0	92.1	94.3	87.7	98.0	92.8	86.7	100.0	92.6
RWA [37]	**93.8**	**98.9**	**97.8**	95.3	**99.4**	95.9	95.8	93.1	98.4	95.3	**92.4**	99.9	96.3
Ours	92.8	**98.9**	97.0	**96.0**	99.0	**97.0**	**96.5**	**97.0**	**99.5**	**95.5**	91.5	**100.0**	**96.8**

Table 2: A comparative study on Office-CalTech dataset using Resnet-50 features (in %).

Method	MNIST-USPS	USPS-MNIST	MNIST-USPS (full)
MMD†[20]	-	-	81.1
DANN†[9]	77.1	73.0	85.1
DSN†[3]	91.3	-	-
ADDA [35]	89.4	90.1	-
CoGAN [18]	91.2	89.1	95.7
DIFA [40]	92.3	89.7	96.2
[29] for (n=2)	92.1	90.0	93.1
DupGAN [15]	-	-	96.0
CY-CADA [14]	-	-	95.6
Generate to adapt [30]	92.8	**90.8**	95.3
Ours	**93.4**	90.3	**96.4**

Table 3: Performance analysis on the MNIST and USPS pairs. † indicates that those methods employ a few labeled target domain samples in their training, as opposed to ours which only uses source training data (in %).

Method	Accuracy
Source only SVM	50.0
TCA [25]	61.0
ITML [6]	70.0
SA [31]	65.0
GFK [10]	70.0
DANN [9]	73.0
Ours	**77.5**

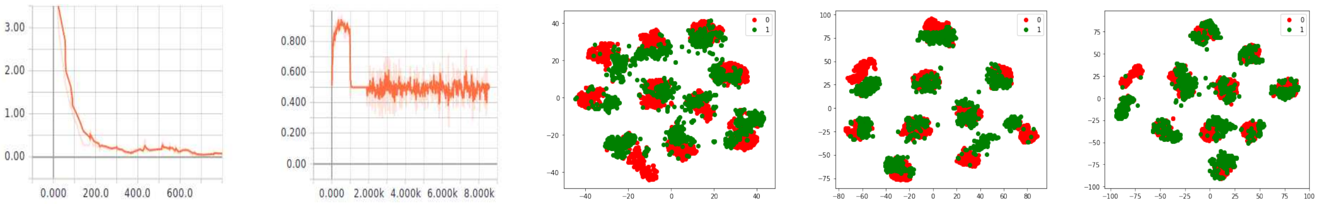
Table 4: A comparative study on the RS dataset (in %).

main classification) (Fig.3(d)), thanks to the class-wise realignment. Further, observe that the training loss gets saturated within 1000 epochs and by then, our improved domain classifier accuracy is fixed around 50%, thus indicating high domain confusion in the learned feature space (Fig.3(a-b)).

Ablation study on the loss: In this regard, we assess the impact of individual loss term (Fig.4). We call the model with all the loss as the *full* model and the model only with source ensemble classifiers (no adaptation) as the *base* model, respectively. The base model provides 76% target accuracy, which is increased by 10% when $\mathcal{L}_D$ is included in *base* loss. Further, the consideration of $\mathcal{L}_T$ in (*base* + $\mathcal{L}_D$) extends this performance significantly by $\geq 5\%$ to produce an overall performance of 92.7%. Finally, the use of the feature regularizer loss $\mathcal{L}_R$ in the *full* model provides the best target accuracy of 93.4%. This clearly indicates the importance $\mathcal{L}_T$ in better aligning the domains with class level information. This trend can be observed for all the dataset. Further note that the $\mathcal{L}_R$ loss induces an enhancement of $1 - 1.5\%$ for all the cases of Office-Home and Office-CalTech dataset.

Effect of the number of classifiers in the ensemble: We analyze the effect of the size of the ensemble by considering 2 to 4 member classifiers in the committee in our *full* model (Fig.4). Note that, while two classifiers are considered, we observe the performance of the ensemble being 93.4%, which is almost 1% more than the performance of the individual classifiers. We find that the performance increases upto 94.4% when the number of classifiers are increased to three and four, respectively. Similar trends can be observed mainly for the large-scale dataset (Office-Home and MNIST-USPS) where $1 - 2\%$ increase in accuracy can be found with more number of classifiers in the ensemble, while the classification performance does not change much for the rather small-scale data (Office-CalTech and the RS data).

Size of $\mathbf{X}^S$ and X^T_{HO}: In this regard, we gradually reduce the size of source training data and carry out the experiments (Fig.4). Note that the α term Eq.7 inversely depends on the number of *i.i.d* samples used during training. As



(a) (b) (c) (d) (e)

Figure 3: (a) Training loss, (b) Accuracy of improved domain classifier, (c) t-SNE of original MNIST, USPS, (d) t-SNE for the domains trained without $\mathcal{L}_T$, (e) t-SNE for full model. 0-1 signifies MNIST and USPS, respectively. The domains are denoted by green and red in (c)-(e).

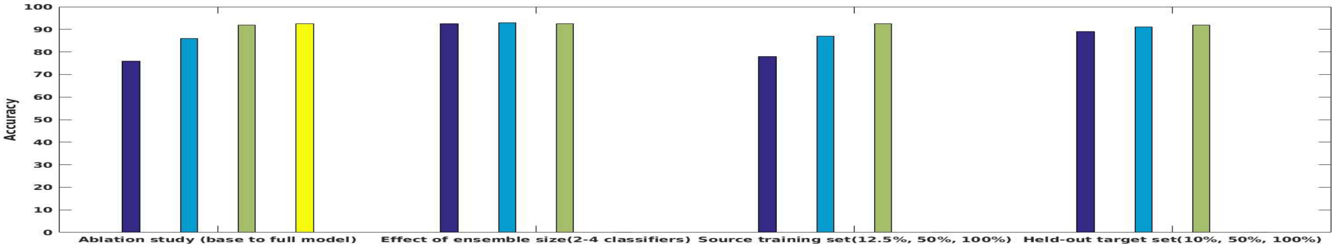


Figure 4: Analysis on **MNIST** $\rightarrow$ **USPS** (P1): From left to right: ablation on loss, ensemble size, size of source training set, size of held-out target subset. Bar colors denote different test scenarios.

expected, the accuracy keeps on decreasing with reduced training set. For instance, we find the accuracy to be 78% when only 250 out of the original 2000 samples are used in $\mathbb{X}^S$ while an accuracy of 87% is obtained when 1000 source domain training samples are deployed. Nonetheless, we consider the entire source domain labeled data as small sized source samples reduces the generalization capability of the method considering the overlapping nature of many of the categories. Similarly, the size of X_{HO}^T matters in effectively training our model. We consider 10%, 50%, and 100% target domain samples separately in this regard. The observation is that the final classification performance increases with more target samples in the held-out set, but not significantly: an increment of $2-3\%$ between the 10% and 100% case (Figure 4).

Sensitivity to λ_1: Given that we find λ_1 influences most the final classification performance among all the individual loss terms (Equation 5 and 6), we report the classification performance for different values of λ_1 in the range $[1, 0.0001]$. We find that the optimal value of λ_1 differs for different dataset. For **MNIST** $\rightarrow$ **USPS** (P1), we obtain the best classification performance for $\lambda_1 = 0.005$. Table 5 depicts the study.

Compactness of the target domain samples: Since the source domain classifiers focus on producing similar class distributions scores for the target samples, this, in turn, makes the target domain samples highly compact. In order to validate this claim, we find the average class compactness over all the categories in terms of the average pairwise distances among all the samples in each class before and after the adaptation. Particularly for **MNIST** $\rightarrow$ **USPS** (P1), we find that the compactness score reduces from 115 to 12, which confirms that high classwise density of the shared space. Similar trend can also be observed for all the dataset.

λ_1	1	0.5	0.05	0.005	0.0001
Accuracy (in %)	90.5	90.4	91.2	93.4	92.4

Table 5: Sensitivity analysis of λ_1 for **MNIST** $\rightarrow$ **USPS**.

5. Conclusions

We propose a novel adversarial loss measure for unsupervised deep DA. In short, our framework addresses two shortcomings of the traditional domain classifier based DA: sub-optimal domain alignment and discrminativeness. In this regard, we find that the enforcement of an additional semantic consistency constraint on the target samples by a classifiers ensemble on source improves the performance of the domain classifier. Besides, we deploy the notion of ensemble learning in our framework and extensive experimental validation confirms the robustness of the approach.

Acknowledgement: B. Banerjee gratefully acknowledges the grant: ECR/2017/000365 by SERB, DST.

References

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

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 7, 2017.

[4] Lorenzo Bruzzone and Claudio Persello. A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability. *IEEE transactions on geoscience and remote sensing*, 47(9):3180–3191, 2009.

[5] Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. Marginalized denoising autoencoders for domain adaptation. *arXiv preprint arXiv:1206.4683*, 2012.

[6] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

[7] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013.

[8] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 1180–1189. JMLR. org, 2015.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[10] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2066–2073. IEEE, 2012.

[11] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 999–1006. IEEE, 2011.

[12] Jisoo Ham, Yangchi Chen, Melba M Crawford, and Joydeep Ghosh. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):492–501, 2005.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[15] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Duplex generative adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1498–1507, 2018.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016.

[19] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016.

[20] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

[21] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1647–1657, 2018.

[22] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2208–2217. JMLR. org, 2017.

[23] Jeroen Manders, Elena Marchiori, and Twan van Laarhoven. Simple domain adaptation with class prediction uncertainty alignment. *arXiv preprint arXiv:1804.04448*, 2018.

[24] Jie Ni, Qiang Qiu, and Rama Chellappa. Subspace interpolation via dictionary learning for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 692–699, 2013.

[25] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[26] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[27] Piyush Rai, Avishek Saha, Hal Daumé III, and Suresh Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32. Association for Computational Linguistics, 2010.

[28] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[29] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. *arXiv preprint arXiv:1712.02560*, 2017.

[30] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018.

[31] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. In *Domain Adaptation in Computer Vision Applications*, pages 153–171. Springer, 2017.

[32] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.

[33] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.

[34] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE geoscience and remote sensing magazine*, 4(2):41–57, 2016.

[35] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, 2017.

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

[37] Twan van Laarhoven and Elena Marchiori. Unsupervised domain adaptation with random walks on target labelings. *arXiv preprint arXiv:1706.05335*, 2017.

[38] Vladimir Naumovich Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[39] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

[40] Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino. Adversarial feature augmentation for unsupervised domain adaptation. *arXiv preprint arXiv:1711.08561*, 2017.

[41] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 2018.

[42] Yifei Wang, Wen Li, Dengxin Dai, and Luc Van Gool. Deep domain adaptation by geodesic distance minimization. In *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*, pages 2651–2657. IEEE, 2017.

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2242–2251. IEEE, 2017.