UNIVERSITY OF CALIFORNIA SAN DIEGO

Instance Level Affinity-Based Transfer for Unsupervised Domain Adaptation

A thesis submitted in partial satisfaction of the
requirements for the degree of Master of Science

in

Computer Science

by

Astuti Sharma

Committee in charge:

Professor Manmohan Chandraker, Chair
Professor Lawrence Saul
Professor Hao Su

2021

The Thesis of Astuti Sharma is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my family and my teachers.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am grateful to Prof. Manmohan Chandraker, for his guidance during my Master's studies. His unique style of advising and ideation were especially helpful to me in developing confidence over my own research ideas and methodology. Without his unwavering support, this thesis would not have been possible.

I owe special thanks to Tarun Kalluri who has been an amazing collaborator and a friend. It's hard to imagine getting through the uncertainties of research without his support.

Most importantly, I would like to thank my family, without whose sacrifices and immense emotional support, this work would not have been possible.

This thesis, in full, has been submitted for publication of the material as it will appear in a conference, 2021, Astuti Sharma, Tarun Kalluri, Manmohan Chandraker. The thesis author was the primary investigator and the first author of this paper.

ABSTRACT OF THE THESIS

Instance Level Affinity-Based Transfer for Unsupervised Domain Adaptation

by

Astuti Sharma

Master of Science in Computer Science

University of California San Diego, 2021

Professor Manmohan Chandraker, Chair

Domain adaptation deals with training models using large scale labeled data from a specific source domain and then adapting the knowledge to certain target domains that have few or no labels. Many prior works learn domain agnostic feature representations for this purpose using a global distribution alignment objective which does not take into account the finer class specific structure in the source and target domains. We address this issue in our work and propose an instance affinity based criterion for source to target transfer during adaptation, called ILA-DA. We first propose a reliable and efficient method to extract similar and dissimilar samples across source and target, and utilize a multi-sample contrastive loss to drive the domain alignment process. ILA-DA simultaneously accounts for intra-class clustering as well as inter-

class separation among the categories, resulting in less noisy classifier boundaries, improved transferability and increased accuracy. We verify the effectiveness of ILA-DA by observing consistent improvements in accuracy over popular domain adaptation approaches on a variety of benchmark datasets and provide insights into the proposed alignment approach.

Chapter 1

Introduction

As stated in [1], in machine learning, if the training data is an unbiased sample of an underlying distribution, then the learned model will make accurate predictions for new samples. However, a domain shift or distributional shift will be observed if the training data is not an unbiased sample. Domain adaptation is concerned with accounting for these types of changes. In other words, domain adaptation allows us to train a model on source domain(s) and apply on a target domain. In unsupervised domain adaptation settings, source domain is labeled whereas the target domain is completely unlabeled.

An example of domain adaptation can be observed in self-driving car systems. These systems may use perception models to identify drivable areas from images. Such models can be trained well from data of a particular place, e.g. San Francisco. However, if we use the same model for a different place, e.g. Bangalore, India, it will not perform well as it has never seen it before. Here the domain adaptation comes to rescue by transferring knowledge from one domain to another.

In this work, we propose a method to leverage instance wise similarities across datasets, called ILA-DA, to improve unsupervised domain adaptation. It is well known that models trained on a large-scale labeled dataset are generally sensitive to domain shifts and do not generalize well to data that lies outside the training distribution [2]. Unsupervised domain adaptation [3, 4, 5] emerged as a feasible alternative to transfer knowledge from a labeled source

domain to one or more unlabeled target domains by minimizing some notion of divergence between the domains [6, 7, 8, 9, 10, 11]. A majority of successful approaches rely on global distribution alignment using adversarial learning [9, 10, 11, 12, 13], where the objective is to learn features that are good enough to fool a discriminator into classifying source samples as target and vice versa. A major limitation with these methods is that while learning domain agnostic feature representations, they do not consider the finer class specific structure of the samples during the alignment resulting in noisy predictions near classifier boundaries. They do not take into account, for example, the fact that the affinity of different categories across source and target towards alignment can be different, which might lead to misalignment of few categories as shown in Fig 1.1. This problem is alleviated to an extent by many follow-up works that make use of target pseudo labels to guide class specific alignment [14, 15, 16, 17, 18, 19, 20]. However, the performance of these approaches is in most cases tied to the reliability of predicted pseudo labels which can be noisy without adequate filtering measures, leading to negative alignment between unrelated categories.

In this work, we address these limitations by proposing a novel adaptation approach called ILA-DA (Instance Level Affinity-based Domain Adaptation). We combine ideas from metric learning literature [21, 22, 23, 24, 24, 25, 26] to perform cross domain transfer by using instance affinity relations between the source and target samples. As opposed to prior works that perform domain level or class level alignment, we show that a much finer knowledge in the form of sample level similarity can be successfully exploited to improve the adaptation process. The main challenge with this approach is that the target domain is completely unlabeled to extract similarity. To overcome this, we propose a nearest neighbor based technique to first construct a pairwise affinity matrix. We then use this knowledge of cross domain positive and negative relations in a multi-sample contrastive learning (MSC) loss that uses multiple positives and negatives across domains in a contrastive learning framework [26, 27].

We identify two advantages using ILA-DA. Firstly, the pairwise similarities provide a relatively stronger signal during training and are shown to be more robust to label corruptions

compared to category predictions in many cases [28, 29]. Secondly, our multi-sample contrastive loss aims to cluster similar samples from across domain closer together while pushing dissimilar samples away to avoid negative transfer. This is especially useful in adaptation across fine-grained datasets, where the challenge, apart from domain shift, is to additionally acknowledge the large intra-class variation within the categories.

The effectiveness of ILA-DA is reflected by improved adaptation accuracy on popular benchmarks like Digits and Office-31 datasets. We also achieve state-of-the-art results on a challenging adaptation dataset Birds-31 [30] without using complementary information such as label-hierarchies and class structure unlike [30], which indicates the usefulness of our MSC loss in handling wide variety of scenarios. We further perform extensive ablations and analysis on our methodological choices. All code and data for our method and baselines will be publicly released.

In summary, the key highlights of the thesis are:

- We propose a novel adaptation frame work ILA-DA. It uses Multi-Sample Contrastive (MSC) loss to perform instance affinity aware transfer by identifying pairwise similarity relations across source and target domains.

- ILA-DA is designed to be general and can be applied to enhance any existing adversarial adaptation approach. We show experimental results while using it in combination with two popular methods, DANN [9] and CDAN [31], and observe consistent improvements over both the baselines.

- We validate the effectiveness of the proposed approach numerically by applying it on multiple tasks from various challenging benchmark datasets used for domain adaptation like Digits, Office-31 and Birds-31 and observe improved accuracies in all the cases, sometimes outperforming the state-of-the-art by a large margin.

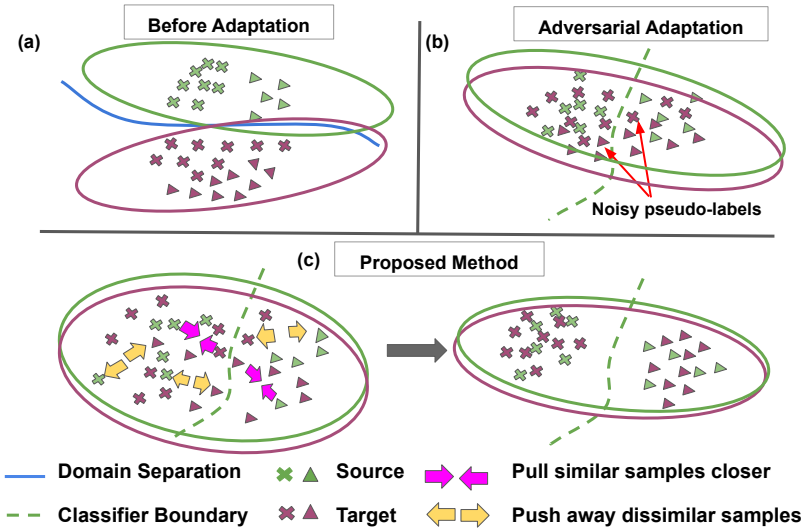


Figure 1.1. Motivation for the proposed approach (**a**), (**b**) Most adversarial learning based adaptation approaches achieve global domain alignment which often leads to misalignment near the classifier boundaries. (**c**) Using our affinity matrix based approach in combination with the proposed MSC loss, we achieve better discrimination between target samples and improve the adaptation.

Chapter 2

Related Work

2.1 Distance/Divergence Based Domain Adaptation

Unsupervised domain adaptation enables training networks on completely unlabeled data by transferring knowledge from a model trained on a different labeled source domain. This is done by minimizing some notion of distance or divergence between the domains [3, 4]. The various notions of divergence include Maximum Mean Discrepancy (also known as MMD) [7, 6, 32, 33, 34, 35, 16, 36, 37, 14] between the feature embeddings of the source and target domains in a RKHS, higher order correlations between the domains [38, 8, 39], optimal transport distance between the source and target [40, 41] and distribution matching using generative [42, 43, 44] or discriminative [9, 10, 11, 12, 13, 31] adversarial learning between a feature generator and a discriminator. As stated in [45], in adversarial domain adaptation, the metric that is used to measure the distance between the source and target domain feature distributions, is crucial for the performance of the domain discriminator and the final adapted model. The adversarial discriminative domain adaptation [10] uses a domain classifier with cross entropy (CE) loss to distinguish the feature vectors from the source domain and the target domain. The minimization of CE between two distributions for domain adaptation is equivalent to the minimization of the Kullback-Leibler (KL) divergence between the two distributions [46]. In this work, we propose complementary improvements to adversarial methods.

2.2 Class-Specific Adaptation

Most of the above works aim to learn domain agnostic feature representations from the source and target data by aligning their global distributions, so that a source classifier can be used on the target. However this does not guarantee alignment between the respective categories which might lead to negative transfer. Recent works alleviated this problem by taking into account class specific properties during adaptation between the domains [47, 48, 16, 13, 49, 18, 14, 19, 20]. Since the target domain is completely unlabeled, these works rely on training co-regularization networks [48, 50], predicting psuedo-labels [18, 14, 51] or computing prototypical [52] representations of source and target categories [19, 20, 53] to assign target classes during training. This makes the performance of these methods dependent on the pseudo-labeling hypothesis, leading to noisy predictions near the classifier boundaries. This is problematic, for example, in fine grained classification setting where the variation within a class is often large. In contrast, we propose a novel sample level transfer criterion which is robust to noisy psuedo-labeling and improves adaptation. A related work is Contrastive Adaptation Network [14], but it is based on MMD and requires k-means clustering after each iteration to update pseudo labels, whereas our ILA-DA is an adversarial approach that uses the proposed affinity matrix combined with a new MSC loss to explicitly model pairwise interactions.

2.3 Metric Learning

Metric learning learns a representation of the data points in an embedded space such that the similarity of the data points is preserved. Data points belonging to same class get close and dissimilar data points get far away. Metric learning uses different loss functions for this. For example, the contrastive loss try to learn representations by making positive pairs attracted and negative pairs separated. In [21], the authors presented a Mahalanobis distance function for the k-nearest neighbors (kNN) classifier and used triplet loss for bringing exemplars from same class together while separating exemplars from different classes. [54] gave a information-theoretic Mahalanobis distance metric by minimizing the differential relative entropy between two distance

functions.

There have been a number of approaches proposed to learn discriminative boundaries between categories using sample-wise [21, 55, 56, 57, 58, 59, 60] or proxy-based [61, 25, 62, 63] metric losses for tasks like face recognition [23, 22, 24], where the challenge is to concurrently address large intra-class variation as well as small inter-class differences. Our multi-sample contrastive (MSC) loss is built on top of the noise contrastive loss [64] and softmax constrastive loss [27, 65], where we extend it to handle multiple positives and negatives at once to leverage sample level relationships useful for adaptation.

2.4 Metric Learning for UDA

While there have been prior works that propose adaptation algorithms for metric learning [66, 67, 68], there have been very few prior works that study the complementary problem of leveraging principles from metric learning to improve regular domain adaptation. In unsupervised domain adaptation settings it is challenging to determine positive and negative pairs as we do not have access to the labels in target domain. Prior works either use triplet loss [69] requiring complex sampling strategy or do not leverage instance level relations [53]. In our work, we acknowledge the need to address intra-class variance within aligned source and target categories for adaptation, which we achieve by proposing a sample level cross dataset transfer mechanism.

Chapter 3

ILA-DA

We first give a brief overview of adversarial adaptation methods, and then introduce our multi sample contrastive (MSC) loss for adaptation followed by construction of affinity matrix.

3.1 Overview of Adversarial Domain Adaptation

In the problem of unsupervised domain adaptation, we have a labeled source dataset $\mathscr{D}^s$:$\{x_i^s, y_i\}_{i=1}^{|\mathscr{D}^s|}$, where $\mathscr{D}^s \sim P_s$ along with an unlabeled target domain $\mathscr{D}^t$:$\{x_i^t\}_{i=1}^{|\mathscr{D}^t|}$ where $\mathscr{D}^t \sim P_t$, and $P_s \neq P_t$. The task is to train a model using these data to make predictions on $\mathscr{D}^t$. We present the overview of the architecture used for training in Fig 3.1. The feature extractor $\mathscr{G}$, which is shared between the source and the target images, extracts the lower dimensional feature representations corresponding to the inputs, given by $f = \mathscr{G}(x)$. The classifier $\mathscr{C}$ then outputs a softmax prediction distribution over the classes, and it is trained using a cross entropy (CE) loss on the labeled source data given by

$$\mathscr{L}_{sup} = \mathbb{E}_{(x,y)\sim\mathscr{D}^s}[-\log[\mathscr{C}(\mathscr{G}(x))]_y], \tag{3.1}$$

where y is the ground truth label corresponding to the source input x and the expectation is taken over all the source data $\mathscr{D}^s$. However, since $P_s \neq P_t$, the classifier trained on source data does not transfer well to target samples, and an adversarial learning strategy [9, 10] is used to alleviate this issue. A domain discriminator $\mathscr{D}$ is trained using $\mathscr{L}_D$ to classify between source and target,

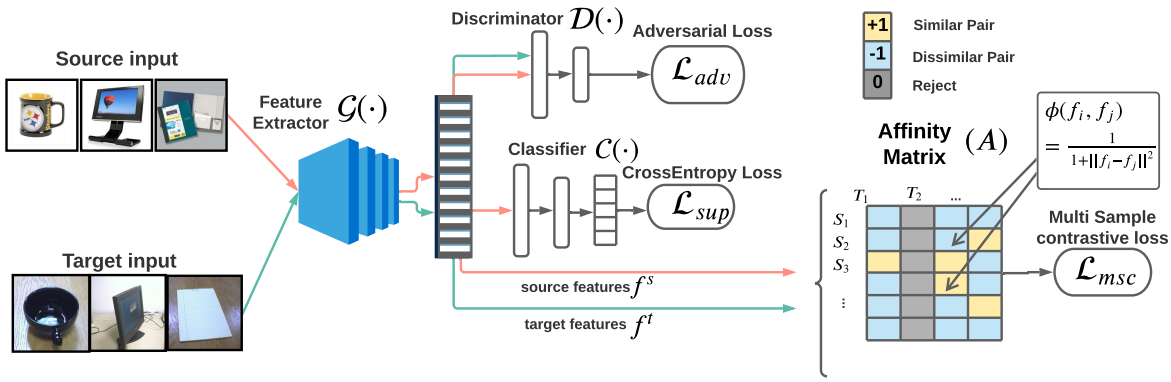


Figure 3.1. Illustration of proposed ILA-DA approach with MSC Loss. Our architecture consists of a feature extractor $\mathscr{G}(.)$ that is shared across source and target domains. The classifier $\mathscr{C}(.)$ is trained to classify the source images using cross entropy loss $\mathscr{L}_{sup}$, while the domain discriminator $\mathscr{D}(.)$ performs domain alignment using adversarial loss $\mathscr{L}_{adv}$. Additionally, we use source and target features to construct an affinity matrix A that holds similarity and dissimilarity relations between the samples (Sec 3.3). We then use this information to cluster categories closer to each other using our proposed multi-sample contrastive loss (Sec 3.2).

while $\mathscr{G}$ is simultaneously trained using $\mathscr{L}_{adv}$ to generate features that confuse the discriminator:

$$\mathscr{L}_{adv} = \mathbb{E}_{x\sim\mathscr{D}^t}[-\log \mathscr{D}(\mathscr{G}(x))], \tag{3.2}$$

$$\mathscr{L}_{D} = -\mathbb{E}_{x\sim\mathscr{D}^s}[\log \mathscr{D}(\mathscr{G}(x))] - \mathbb{E}_{x\sim\mathscr{D}^t}[\log(1-\mathscr{D}(\mathscr{G}(x)))]. \tag{3.3}$$

Min-max training between $\mathscr{L}_D$ and $\mathscr{L}_{adv}$ then yields domain invariant features. However, this is not enough to guarantee class specific alignment between source and target, so we present our proposed affinity matrix based adaptation next.

3.2 Multi-Sample Contrastive (MSC) Loss

To enforce the class-level alignment constraint, we first find the sample level similarity scores among the source and target samples in a mini-batch and use them in our multi-sample contrastive (MSC) loss. However, we do not have labels for the target domain, so we follow a kNN based approach to assign each target sample in a mini-batch with a label belonging to nearby source samples. We then construct an affinity matrix A, in which $A_{ij} = 1$ if the i^{th} source sample

from the mini-batch is similar to j^{th} target sample from the mini-batch, $\mathrm{A}_{ij} = -1$ if it is not. This is explained in detail in Sec 3.3. Assuming we have constructed such an affinity matrix A, we use this information to construct positive and negative samples corresponding to a source sample x_i. Specifically, let B_S and B_T be the source and target batches respectively. Then, for each source sample $x_i \in B_s$, we identify the set of positive target pairs as $B_T^{i+} = \{x_j \in B_T | \mathrm{A}_{ij} = 1\}$, and negative pairs as $B_T^{i-} = \{x_j \in B_T | \mathrm{A}_{ij} = -1\}$. We then use this information to pull similar samples across source and target closer to each other, while pushing away dissimilar samples using our MSC loss given by:

$$\mathscr{L}_{MSC}^{i} = -\log \frac{\sum\limits_{j \in B_T^{i+}} e^{\phi(f_i, f_j)}}{\sum\limits_{j \in B_T^{i+}} e^{\phi(f_i, f_j)} + \sum\limits_{j \in B_T^{i-}} e^{\phi(f_i, f_j)}}, \tag{3.4}$$

where B_S and B_T denote the source and target batches respectively, f are the features computed as the output of $\mathscr{G}(x)$ and $\phi(.,.)$ is any metric that takes the features and outputs a similarity score. The overall loss is computed as the average across all the source samples from the mini-batch B_S:

$$\mathscr{L}_{MSC} = \frac{1}{|B_S|} \sum_{i \in B_S} \mathscr{L}_{MSC}^{i}. \tag{3.5}$$

Empirically, we observe best results when using normalized inverse Euclidean distance [70] as the similarity metric ϕ:

$$\phi(f_i, f_j) = \frac{1}{1 + ||f_i - f_j||^2}. \tag{3.6}$$

This process is illustrated in Fig 3.2. Similar kind of contrastive loss is used for learning representations from unlabeled image and video in [27, 71, 72, 73] where positives come from transformed versions of inputs unlike ILA-DA. Furthermore, contrastive loss is shown to work well for large intra-class variations empirically in [72, 74] and theoretically in [75]. ILA-DA demonstrates similar benefits, while additionally accounting for possible domain gap between the positive and negative pairs. From (3.4), we can observe that if $\mathrm{A}_{ij} = 1$, indicating similar

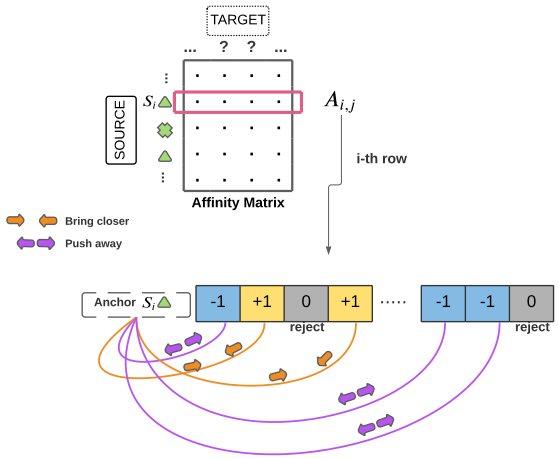


Figure 3.2. Illustration of our MSC loss. i^{th} row of affinity matrix A_i contains similarity information of i^{th} source sample with every target sample. MSC loss uses these relations to attract positive sample from target while separating negative ones.

pairs, then the similarity metric needs to be higher to minimize the loss. Likewise, if $A_{ij} = -1$, then the similarity score would be driven down to zero, as we require it to be. We now explain in detail the method to construct the affinity matrix A.

3.3 Constructing the Affinity Matrix

Recall that the target dataset is completely unlabeled, so obtaining similarity scores is not trivial. Using source classifier to assign pseudo labels is an option, but it would be noisy during initial stages of training and empirically suboptimal (Sec 4.3, Tab 4.5). Instead, we rely on a k-nearest neighbor approach followed by a ratio test to assign confident target labels. For every target sample $x_j \in B_T$, we take the k nearest neighbors ranked using the same similarity metric $\phi(f_j, .)$ from the source mini-batch. Then, the target sample is assigned to the class that is most common among its source neighbors and we populate the j^{th} column of the affinity matrix A using this assignment. That is,

$$A_{ij} = \begin{cases} 1, & \text{if } y_i = \hat{y}_j \\ -1, & \text{if } y_i \neq \hat{y}_j. \end{cases}$$

Although the similarity and dissimilarity relations can now be directly read off the affinity matrix A, we did not yet account for the fact that some of the psuedo labels can be noisy. After constructing the affinity matrix A, we filter out possible noisy pseudo-labels. For this, we use a rejection based confidence measure commonly used in kNN literature based on a neighborhood similarity ratio test [76, 77]. Denote using N_j^l the set of like samples from source in the neighborhood of a target sample $x_j \in B_T$, given by $N_j^l : \{x_i \in B_S | y_i = \hat{y}_j\}$. Similarly, the set of unlike samples from source in the neighborhood is given by N_j^u, where $N_j^u : \{x_i \in B_S | y_i \neq \hat{y}_j\}$. We calculate the confidence score of a particular pseudo label prediction Γ_j using the ratio of aggregate similarity between the sample and the like and unlike sets. That is,

$$\Gamma_j = \frac{\sum_{x_i \in N_j^l} \phi(f_j, f_i)}{\sum_{x_i \in N_j^u} \phi(f_j, f_i)}. \tag{3.7}$$

We then choose a sampling factor μ, and select the top μ fraction of target samples and declare them to be confident, and for the rest of target samples, we put $A_{ij} = 0$ and do not use them anymore in the MSC loss Eq (3.4). For example, if sampling factor $\mu = .75$ with a batch size of 128, we select the top 96 target samples ranked based on their prediction confidence Γ. Since the number of unlike samples are generally much higher than the number of like samples, we only take the top m samples in the summations in Eq (3.7) to balance the aggregate between the like and unlike sets. We chose m to be the maximum possible similar samples across datasets. In our case m is equal to the size of each class in source mini-batch. This way, we will be left with pairwise similarity scores between pairs of source and target samples which pass the similarity ratio test. Further analysis of such a psuedo labeling procedure, including the sensitivity to the sampling factor μ, is presented in Sec 4.3. The complete algorithm is summarized in Algo 1.

Although many prior works have considered a psuedo-labeling criterion for assigning target labels during training [51, 19], the advantage we provide lies in the fact that our MSC loss takes sample level similarities with an explicit push-pull objective which is greatly useful to model finer category separation. Also, we have $O(n^2)$ psuedo-labels in each mini-batch of size

ALGORITHM 1. Instance Affinity Based Adaptation during each iteration.

Require: Class balanced mini-batches for source $B_S \in \mathscr{D}^s$ and randomly sampled target mini-batches $B_T \in \mathscr{D}^t$

Require: Feature extractor $\mathscr{G}(.)$

Require: Similarity metric $\phi(.,.)$

$A_{ij} = 0 \; \forall i \in \{1,2..,|B_S|\}$, $j \in \{1,2..,|B_T|\}$ x_j in B_T ▷ Construct affinity matrix

$\hat{y}_j = kNN(B_S, x_j)$ (Sec 3.3) x_i in B_S

$A_{ij} = 1$ **if** $y_i = \hat{y}_j$ **else** $A_{ij} = -1$ x_j in B_T

$\Gamma_j(x_j)$ (Eq 3.7) ▷ Compute Similarity Ratio

$B_T^F = Filter(B_T, \Gamma, \mu)$ ▷ Select confident pseudo-labels using similarity ratio test.

$Loss = MSC(B_S, B_T^F)$ (Eq 3.4) ▷ Compute MSC loss.

n, so we will be left with a strong signal even after removing lesser confident predictions. In contrast to [78], we extract kNN neighbors across source and target, calculate sample-sample as opposed to sample prototype relations for use in our MSC loss.

Finally, when we randomly sample mini batches from source and target, it might so happen that some classes might not get picked in source, which is problematic. For example, some target samples might not have a corresponding true source sample leading to incorrect psuedo labels, or some source sample might get paired with a dissimilar target sample in our MSC loss in Eq (3.4). To avoid this issue, we perform class balanced mini batch sampling only on the source dataset, in which we make sure that all classes have equal representation in all the sampled source mini batches B_S. Unlabeled target mini-batches are sampled randomly.

Chapter 4

Results

In this section, we conduct extensive experiments on multiple domain adaptation benchmarks to verify the effectiveness of ILA-DA approach. We next present the datasets used to evaluate our results, baselines methods we compared against, followed by results and discussion.

4.1 Experimental Details

We investigate the performance of our model on three different kinds of benchmark datasets used for domain adaptation, namely Digits, Office-31 and Birds-31.

4.1.1 Digits

We use SVHN, MNIST and USPS consisting of images of digits $0-9$. We explore the adaptation tasks between **MNIST** $\rightarrow$ **USPS**, **USPS** $\rightarrow$ **MNIST** and **SVHN** $\rightarrow$ **MNIST**.

4.1.2 Office-31

This setting consists of images from 31 categories from three different domains, namely Amazon (**A**), Webcam (**W**) and DSLR (**D**). We show results for all the 6 task pairs **A** $\rightarrow$ **W, D** $\rightarrow$ **W, W** $\rightarrow$ **D, A** $\rightarrow$ **D, D** $\rightarrow$ **A** and **W** $\rightarrow$ **A**. Following prior works, we report results on the complete unlabeled examples of the target domain.

4.1.3 Birds-31

This dataset is recently proposed by [30] for fine grained adaptation consisting of different types of birds. We use it to verify our argument that our MSC loss performs efficiently even with datasets that possess large intra-class and small inter-class variation. It consists of three domains, namely, 1848 images from CUB-200-2011 (**C**) [79], 2988 images from NABirds (**N**) [80] and 2857 images from iNaturalist2017 (**I**) datasets from the 31 common classes among the three. We show the adaptation results on six transfer tasks formed from three domains: **C** $\rightarrow$ **I, I** $\rightarrow$ **C, I** $\rightarrow$ **N, N** $\rightarrow$ **I, C** $\rightarrow$ **N** and **N** $\rightarrow$ **C**.

4.1.4 Training details

Following prior works [31, 53], we use LeNet architecture for digits and use ResNet-50 (pretrained on Imagenet) as the feature extractor $\mathscr{G}$ for the Office-31 and Birds-31 datasets, while the classifier $\mathscr{C}$ is made up of fully connected layers. For achieving training stability, we observe that it is essential to pretrain the model on the labeled source dataset for a few iterations before introducing our constrastive loss.

We use mini-batch SGD with a learning rate of 0.001 for Office and 0.03 for birds. For the classifier we multiply the learning rate by 10. We use a similar annealing strategy as used in [9]. Further details on the hyperparameter settings are presented in the supplementary material.

To illustrate the benefits of the proposed MSC loss, we employ it on top of two competing adaptation benchmarks in DANN [9] and CDAN [31], while noting that our loss is general and applicable in combination with any adversarial adaptation approach. For experiments with DANN, we replace the adversarial loss with a gradient reversal layer.

4.1.5 Baselines

We focus our comparison against works which use adversarial learning strategy to perform global domain level alignment such as **DAN** [7], **RTN** [81], **ADDA** [10], **GTA** [43],

Table 4.1. Accuracy (%) on Digits for unsupervised domain adaptation. Results shown for a value of $k = 3$ and $\mu = 0.75$.

Method	M → U	U → M	S → M	Avg.
Source Only	76.7	63.4	67.1	69.1
DANN [9]	90.8	93.95	83.11	89.29
ADDA [10]	89.4	90.1	76.0	85.2
DSN [11]	91.3	-	82.7	-
ATT [48]	-	-	85.0	-
ILA-DA (with DANN)	92.43	97.32	91.84	93.83
CDAN [31]	93.9	96.9	88.5	93.1
ILA-DA (with CDAN)	**94.87**	**97.47**	**92.30**	**94.88**

DAA [82] and **CDAN** [31] as well as works which perform class aware alignment such as **MCD** [83], **SimNet** [53], **MADA** [17]. For Birds-31, we additionally verify our result with prior fine grained adaptation work, **PAN** [30]. Finally, we have **ILA + DANN**, which is using ILA-DA approach on top of DANN and **ILA + CDAN** which uses ILA-DA in combination with CDAN. We compare the task-wise accuracies and also report the average accuracies across all the transfer tasks. Our training and evaluation scripts are publicly released.

4.2 Comparison with baselines

4.2.1 Digits

In Tab 4.1, we show the results for adaptation using our method with MSC loss. We observe that we outperform prior methods by a significant margin when we use CDAN in combination with ILA-DA. On **MNIST → USPS** we observe an improvement from 90.8 to 92.43 while using ILA-DA + DANN and 93.9 to 94.87 with ILA-DA + CDAN, indicating the usefulness of our MSC loss for improving existing methods for domain adaptation. Similar improvements can be observed for all other dataset settings as well, for instance, accuracy goes up from 88.5 to 92.30 in the case of **SVHN → MNIST** using ILA-DA with CDAN.

4.2.2 Office-31

We present results on the 6 transfer tasks on Office-31, including their average, in Tab 4.2. We observe that we achieve an accuracy of 89.30% on the average, outperforming all the competing baselines, which includes prior works that perform global domain alignment [81, 10],

Table 4.2. Results for domain adaptation on Office-31 adaptation setting using Resnet-50 for 6 transfer tasks among three domains: Amazon (A), Webcam (W) and Dslr (D). Our method shows consistent improvements. All the baselines as well as ours use ResNet-50 as the backbone architecture. Results shown for $k = 5$ and $\mu = 0.67$.

Method	A → W	D → W	W → D	A → D	D → A	W → A	Avg.
ResNet-50	68.4	96.7	99.3	68.90	62.50	60.70	76.1
DAN [7]	80.5	97.1	99.6	78.6	63.6	62.8	80.4
RTN [81]	84.5	96.8	99.4	77.5	66.2	64.8	81.6
DANN [9]	82.0	96.9	99.1	79.7	68.2	67.4	82.2
ADDA [10]	86.2	96.2	98.4	77.8	69.5	68.9	82.9
MCD [83]	88.6	98.5	100.0	92.2	69.5	69.7	86.5
SimNet [53]	88.6	98.2	99.7	85.3	73.4	71.8	86.2
GTA [43]	89.5	97.9	99.8	87.7	72.8	71.4	86.5
CDAN [31]	93.1	98.2	**100.0**	89.8	70.1	68.0	86.6
CDAN+E [31]	94.1	98.6	**100.0**	92.9	71.0	69.3	87.7
DAA [82]	86.8	99.3	**100.0**	88.8	**74.3**	73.9	87.2
SAFN [84]	88.8	98.4	99.8	87.7	69.8	69.7	85.7
MADA [17]	90.0	97.4	99.6	87.8	70.3	66.4	85.2
ILA-DA (with DANN)	89.05	98.49	**100.0**	86.55	69.47	69.72	85.54
ILA-DA (with CDAN)	**95.72**	**99.25**	**100.0**	**93.37**	72.10	**75.40**	**89.30**

as well as those that model finer class separation [83, 53, 17] like us, highlighting the advantages of our MSC loss in comparison to competing approaches. Finally, to testify that our loss is generally applicable, we show that it improves accuracy over both the approaches DANN [9] and CDAN [31], consistently over all the tasks (by 3.3% and 2.7% on average, respectively). This result underlines the necessity for our sample aware class-specific transfer in addition to global domain alignment.

4.2.3 Birds-31

The difficulty in this setting lies in the fact that birds from same class but different domains look quite distinct, sometimes more different than images from another class. We verify the results on all 6 transfer tasks on Birds-31 dataset in Tab 4.3, and show that ILA-DA outperforms prior works across all the tasks. Due to the intra-class variation in the dataset and small inter-class distances, prior works that rely on global alignment objectives [7, 6, 10] do not perform any better than a source-only model (*ResNet-50* baseline), possibly because they suffer from negative alignment. However, our MSC loss explicitly accounts for the instance level relations to model category separation, which pulls similar samples from both datasets closer while pushing away dissimilar ones. As a result, we improve the accuracy over DANN on all

Table 4.3. Results for domain adaptation on fine-frained adaptation setting, shown for 3 challenging datasets, namely CUB-200-2011 (C), iNaturalist2017 (I) and NABirds (N). We perform consistently better than all other methods by explicitly modeling the finegrained nature of the adaptation process. All the methods use ResNet-50 pretrained on ImageNet. All the baseline numbers taken from [30]. Results shown for $k = 3$ and $\mu = 0.33$.

Method	C → I	I → C	I → N	N → I	C → N	N → C	Avg.
ResNet-50	64.25	87.19	82.46	71.08	79.92	89.96	79.14
DAN [7]	63.9	85.86	82.91	70.67	80.64	89.40	78.90
DANN [9]	64.59	85.64	80.53	71.00	79.37	89.53	78.44
JAN [6]	63.69	86.29	83.34	71.09	81.06	89.55	79.17
ADDA [10]	63.03	87.26	84.36	72.39	79.69	89.28	79.33
MADA [17]	62.03	89.99	87.05	70.99	81.36	92.09	80.50
MCD [83]	66.43	88.02	85.57	73.06	82.37	90.99	81.07
CDAN [31]	68.67	89.74	86.17	73.80	83.18	91.56	82.18
SAFN [84]	65.23	90.18	84.71	73.00	81.65	91.47	81.08
PAN [30]	69.79	90.46	88.10	75.03	84.19	92.51	83.34
ILA-DA (with DANN)	69.55	93.13	87.15	74.69	83.40	93.89	83.63
ILA-DA (with CDAN)	**72.77**	**93.83**	**90.36**	**78.09**	**86.58**	**94.53**	**86.03**

the tasks, and average accuracy from 78.44% to 83.63%. In fact, with an average accuracy of 86.03% we achieve the *new state-of-the-art result* using ILA-DA in combination with CDAN. More remarkably, ILA-DA+CDAN even outperform PAN [30], that is specifically designed for fine-grained adaptation by roughly 3% without demanding access to any label structure and class hierarchy during training unlike [30], which highlights the usefulness of modeling instance level loss for challenging adaptation problems.

4.3 More Qualitative Results

4.3.1 Importance of MSC Loss

We testify the effectiveness of the proposed multisample contrastive loss in modeling the instance level relations by comparing it to another commonly used metric loss, namely triplet loss. We replace the loss used in Eq (3.4) by triplet loss, by deriving positives and negatives from the affinity matrix. We use similarity metric $\phi(.)$, and choose the nearest negative sample and farthest positive sample as hard negative and hard positive respectively. From Tab 4.4, we first observe that both triplet loss as well as MSC loss improve over CDAN baseline, which indicates the usefulness of adding metric learning losses over adversarial methods for better alignment. Further, we also observe that replacing MSC loss by triplet loss leads to drop in

Table 4.4. Comparison of triplet Loss vs. MSC Loss for metric learning. Results shown for A→D and W→A tasks and avg. of all 6 tasks from Office-31 dataset.

Method	A→D	W→A	Avg.
CDAN[31]	89.8	68.0	86.6
Triplet + CDAN	90.20	73.94	87.63
MSC + CDAN	**93.37**	**75.40**	**89.30**

Table 4.5. Comparison of kNN vs. Classifier based psuedo-labeling schemes. Results shown for A→D and W→A tasks and avg. of all 6 tasks from Office-31 dataset for $k = 5$, μ=0.67.

Method	A→D	W→A	Avg.
CDAN[31]	89.8	68.0	86.6
classifier based	88.35	70.11	86.68
kNN based	**93.37**	**75.40**	**89.30**

accuracy from 93.37% to 90.20% on A→D and from 75.40% to 73.94% on W→A settings on Office-31 dataset. From this, we conclude that for improving domain adaptation, modeling multiple instance relations at once using MSC loss is simpler and more powerful than triplet loss.

4.3.2 Choice of psuedo-labeling

In proposed ILA-DA, the psuedo labeling process for the target examples is driven by finding the k nearest source neighbors in the feature space. Alternatively, we can directly use the source classifier predictions as psuedo labels [19, 17]. To tease out the differences between these alternatives, we compare against such a classifier based psuedo labeling method which filters the target samples using softmax scores as an indicator for the prediction confidence, in Tab 4.5 . We observe that our kNN based approach provides significant benefit over the classifier based counterpart on all the tasks, with a 2.62% boost in accuracy on average.

4.3.3 Effect of sampling factor

We investigate the effect of the sampling parameter μ, used to threshold the similarity ratio Γ in Eq (3.7). Intuitively, a very high value of μ would lead to many noisy psuedo labels being accepted leading to poor optimization, while a low value would eliminate even moderately

Table 4.6. Effect of sampling fraction μ. Results shown for A→D and W→A tasks and avg. of all 6 tasks from Office-31 dataset for $k = 5$.

Method	A→D	W→A	Avg.
CDAN[31]	89.8	68.0	86.6
ILA-DA , μ=0.33	90.75	71.95	87.91
ILA-DA , μ=0.50	91.95	74.33	88.53
ILA-DA , μ=0.67	**93.37**	**75.40**	**89.30**
ILA-DA, μ=1.00	92.33	70.39	87.92

Table 4.7. Effect of number of neighbors k used in psuedo-labeling. Results are shown for A→D and W→A tasks and avg. of all 6 tasks from Office-31 dataset.

Method	A→D	W→A	Avg.
CDAN[31]	89.8	68.0	86.6
ILA-DA, k=1	91.96	69.93	87.46
ILA-DA, k=3	91.16	75.15	88.87
ILA-DA, k=5	**93.37**	**75.40**	**89.30**

confident positives which could be useful training signal. In fact, from Tab 4.6 we observe that a value of $\mu = 0.67$ is optimal, which corresponds to accepting the top two-thirds of the psuedo-label predictions.

4.3.4 Effect of k

We show the effect of k in the kNN process in Tab 4.7. We observe that the average accuracy on Office-31 dataset is highest for $k = 5$. We provide further analysis on the influence of k in supplementary material. In general, we find that a value of $k > 1$ is beneficial for reliable psuedo-labeling, as it helps handle noisy predictions around classifier boundaries.

4.3.5 Visualizing the Affinity Matrix

We visualize the affinity matrix A in Sec 3.3 to get an idea of the reliability of predicted pseudo-labels. For a mini-batch of size 120, we plot the 120×120 affinity matrix A in Fig 4.1, grouped by the class ordering. Here, (a) is the affinity matrix constructed using the ground truth similarities. We observe that the unfiltered affinity matrix in (b) already does a good job in accurately predicting the similarity (red , +1) and dissimilarity(yellow , -1) relations between source and target. Furthermore, we filter out noisy pseudo labels using our filtering approach

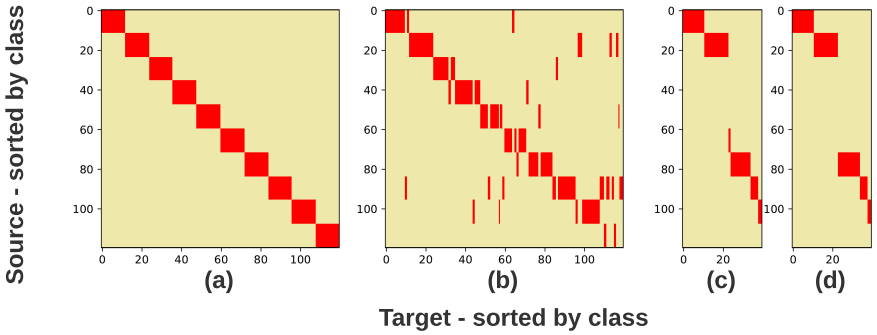
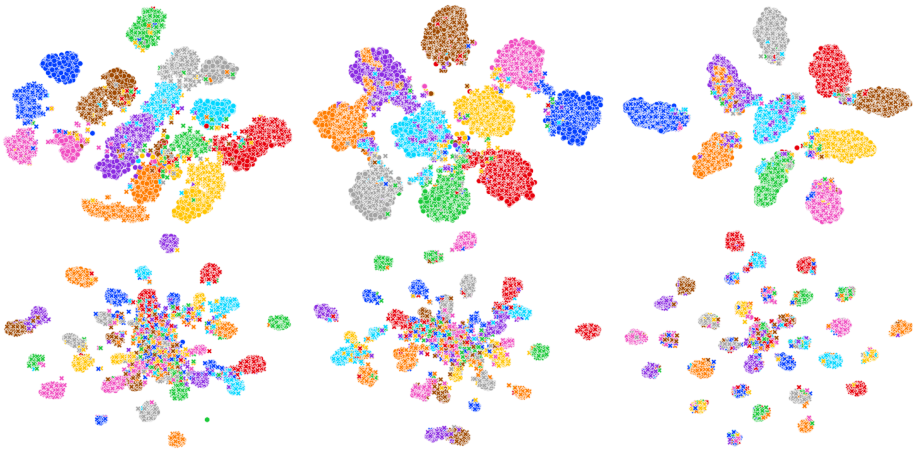


Figure 4.1. Visualization of Affinity matrix A consisting of similarity relations between the source and target for a subset of samples. The computed affinity matrix (b) is close to the ground truth affinity matrix (a), and we further close the gap by efficiently filtering wrong predictions (ground truth in (c) and (d)). Results shown for task M→U from Digits at 40-th epoch during training.

discussed in Sec 3.3, and find that the affinity matrix after filtering (shown in (d)) is much more closer to ground truth affinity matrix, in (c), which verifies the robustness of our pseudo labeling approach.

4.3.6 Feature visualization

We provide the tSNE visualization of the learned features for two different dataset settings in Fig 4.2. On both these settings, we observe better domain alignment as well as target category separation using ILA-DA. Note that although DANN does succeed in aligning the source and target domains, it does not necessarily produce discriminative features, which is addressed by ILA-DA.



(a) No Adapt **(b)** DANN **(c)** ILA-DA

Figure 4.2. tSNE visualizations of source and target features belonging to 10 classes from Digits (S→M) (top row) and 31 classes from birds (N→I) (bottom row). Here, 4.2a shows tSNE with no adaptation. While DANN [9] (4.2b) is only successful in domain alignment, our proposed ILA-DA approach additionally improves category separation on target domain (4.2c).

Chapter 5

Conclusion

In this work, we leverage principles from metric learning to improve domain adaptation. We propose an affinity matrix based approach, ILA-DA, that uses a multi sample contrastive loss to explicitly model instance level interactions across source and target. We show that this helps in improving category separation while preventing negative alignment. The proposed approach is general, and can be easily applied on top of any existing adversarial adaptation method. We show numerical results on various challenging benchmark datasets and perform favorably against many existing adaptation methods.

As with any method that extracts pairwise similarities, the process of constructing the affinity matrix at each iteration is memory intensive. Given current limits on memory, our model may handle a reasonable number of categories across source and target. In future work, we aim to devise newer sampling strategies for affinity matrix construction that allow handling much larger number of classes.

Broader Impact

The broader positive impact of our work would be to inspire methods in computer vision and associated industries such as automotives, to label large amount of generated data which otherwise requires significant human intervention for data labelling.

Our code will be publicly released to encourage further research in the community.

This thesis, in full, has been submitted for publication of the material as it will appear in a conference, 2021, Astuti Sharma, Tarun Kalluri, Manmohan Chandraker. The thesis author was the primary investigator and the first author of this paper.

Bibliography

[1] W. M. Kouw and M. Loog, "An introduction to domain adaptation and transfer learning," 2019.

[2] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, pp. 1521–1528, IEEE, 2011.

[3] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," *Advances in neural information processing systems*, vol. 19, pp. 137–144, 2006.

[4] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[5] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European conference on computer vision*, pp. 213–226, Springer, 2010.

[6] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *International conference on machine learning*, pp. 2208–2217, PMLR, 2017.

[7] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*, pp. 97–105, PMLR, 2015.

[8] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*, pp. 443–450, Springer, 2016.

[9] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, pp. 1180–1189, PMLR, 2015.

[10] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176, 2017.

[11] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Advances in neural information processing systems*, pp. 343–351, 2016.

[12] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3722–3731, 2017.

[13] O. Sener, H. O. Song, A. Saxena, and S. Savarese, "Learning transferrable representations for unsupervised domain adaptation," in *Advances in Neural Information Processing Systems*, pp. 2110–2118, 2016.

[14] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4893–4902, 2019.

[15] P. Haeusser, T. Frerix, A. Mordvintsev, and D. Cremers, "Associative domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2765–2773, 2017.

[16] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proceedings of the IEEE international conference on computer vision*, pp. 2200–2207, 2013.

[17] Z. Pei, Z. Cao, M. Long, and J. Wang, "Multi-adversarial domain adaptation," *arXiv preprint arXiv:1809.02176*, 2018.

[18] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2507–2516, 2019.

[19] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in *International Conference on Machine Learning*, pp. 5423–5432, 2018.

[20] Y. Pan, T. Yao, Y. Li, Y. Wang, C.-W. Ngo, and T. Mei, "Transferrable prototypical networks for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2239–2247, 2019.

[21] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification.," *Journal of Machine Learning Research*, vol. 10, no. 2, 2009.

[22] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

[23] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

[24] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

[25] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, and R. Jin, "Softtriple loss: Deep metric learning without triplet sampling," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6450–6458, 2019.

[26] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 1735–1742, IEEE, 2006.

[27] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[28] Y.-C. Hsu and Z. Kira, "Neural network-based clustering using pairwise constraints," *arXiv preprint arXiv:1511.06321*, 2015.

[29] Y.-C. Hsu, Z. Lv, J. Schlosser, P. Odom, and Z. Kira, "Multi-class classification without multi-class labels," *arXiv preprint arXiv:1901.00544*, 2019.

[30] S. Wang, X. Chen, Y. Wang, M. Long, and J. Wang, "Progressive adversarial networks for fine-grained domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9213–9222, 2020.

[31] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

[32] T. Ming Harry Hsu, W. Yu Chen, C.-A. Hou, Y.-H. Hubert Tsai, Y.-R. Yeh, and Y.-C. Frank Wang, "Unsupervised domain adaptation with imbalanced cross-domain data," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4121–4129, 2015.

[33] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2272–2281, 2017.

[34] M. Baktashmotlagh, M. Harandi, and M. Salzmann, "Distribution-matching embedding for visual domain adaptation," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3760–3789, 2016.

[35] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2010.

[36] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[37] E. Zhong, W. Fan, J. Peng, K. Zhang, J. Ren, D. Turaga, and O. Verscheure, "Cross domain distribution adaptation via kernel mapping," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1027–1036, 2009.

[38] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," *arXiv preprint arXiv:1511.05547*, 2015.

[39] P. Morerio, J. Cavazza, and V. Murino, "Minimal-entropy correlation alignment for unsupervised deep domain adaptation," *arXiv preprint arXiv:1711.10288*, 2017.

[40] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 9, pp. 1853–1865, 2016.

[41] B. Bhushan Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty, "Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 447–463, 2018.

[42] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *arXiv preprint arXiv:1611.02200*, 2016.

[43] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8503–8512, 2018.

[44] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*, pp. 1989–1998, PMLR, 2018.

[45] P. Xu, P. Gurram, G. Whipps, and R. Chellappa, "Wasserstein distance based domain adaptation for object detection," *arXiv preprint arXiv:1909.08675*, 2019.

[46] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, pp. 214–223, PMLR, 2017.

[47] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A dasvm classification technique and a circular validation strategy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 5, pp. 770–787, 2009.

[48] K. Saito, Y. Ushiku, and T. Harada, "Asymmetric tri-training for unsupervised domain adaptation," *arXiv preprint arXiv:1702.08400*, 2017.

[49] W. Zhang, W. Ouyang, W. Li, and D. Xu, "Collaborative and adversarial network for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3801–3809, 2018.

[50] A. Kumar, P. Sattigeri, K. Wadhawan, L. Karlinsky, R. Feris, B. Freeman, and G. Wornell, "Co-regularized alignment for unsupervised domain adaptation," in *Advances in Neural Information Processing Systems*, pp. 9345–9356, 2018.

[51] C. Chen, W. Xie, W. Huang, Y. Rong, X. Ding, Y. Huang, T. Xu, and J. Huang, "Progressive feature alignment for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 627–636, 2019.

[52] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, pp. 4077–4087, 2017.

[53] P. O. Pinheiro, "Unsupervised domain adaptation with similarity learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8004–8013, 2018.

[54] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning*, pp. 209–216, 2007.

[55] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[56] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4004–4012, 2016.

[57] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in neural information processing systems*, pp. 1857–1865, 2016.

[58] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5022–5030, 2019.

[59] S. Kim, M. Seo, I. Laptev, M. Cho, and S. Kwak, "Deep metric learning beyond binary supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2288–2297, 2019.

[60] B. Yu and D. Tao, "Deep metric learning with tuplet margin loss," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6490–6499, 2019.

[61] N. Aziere and S. Todorovic, "Ensemble deep manifold similarity learning using hard proxies," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7299–7307, 2019.

[62] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 360–368, 2017.

[63] S. Kim, D. Kim, M. Cho, and S. Kwak, "Proxy anchor loss for deep metric learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3238–3247, 2020.

[64] A. Mnih and K. Kavukcuoglu, "Learning word embeddings efficiently with noise-contrastive estimation," in *Advances in neural information processing systems*, pp. 2265–2273, 2013.

[65] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. Eslami, and A. v. d. Oord, "Data-efficient image recognition with contrastive predictive coding," *arXiv preprint arXiv:1905.09272*, 2019.

[66] K. Sohn, W. Shang, X. Yu, and M. Chandraker, "Unsupervised domain adaptation for distance metric learning," in *International Conference on Learning Representations*, 2018.

[67] B. Geng, D. Tao, and C. Xu, "Daml: Domain adaptation metric learning," *IEEE Transactions on Image Processing*, vol. 20, no. 10, pp. 2980–2989, 2011.

[68] Z. Ding and Y. Fu, "Robust transfer metric learning for image classification," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 660–670, 2016.

[69] I. H. Laradji and R. Babanezhad, "M-adda: Unsupervised domain adaptation with deep metric learning," in *Domain Adaptation for Visual Understanding*, pp. 17–31, Springer, 2020.

[70] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[71] D. Gordon, K. Ehsani, D. Fox, and A. Farhadi, "Watching the world go by: Representation learning from unlabeled videos," *arXiv preprint arXiv:2003.07990*, 2020.

[72] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

[73] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[74] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

[75] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi, "A theoretical analysis of contrastive unsupervised representation learning," *arXiv preprint arXiv:1902.09229*, 2019.

[76] S. J. Delany, P. Cunningham, D. Doyle, and A. Zamolotskikh, "Generating estimates of classification confidence for a case-based spam filter," in *International conference on case-based reasoning*, pp. 177–190, Springer, 2005.

[77] B. V. Dasarathy, "Nearest unlike neighbor (nun): an aid to decision confidence estimation," *Optical Engineering*, vol. 34, no. 9, pp. 2785–2793, 1995.

[78] J. Liang, R. He, Z. Sun, and T. Tan, "Exploring uncertainty in pseudo-label guided unsupervised domain adaptation," *Pattern Recognition*, vol. 96, p. 106996, 2019.

[79] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[80] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604, 2015.

[81] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Advances in neural information processing systems*, pp. 136–144, 2016.

[82] G. Kang, L. Zheng, Y. Yan, and Y. Yang, "Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 401–416, 2018.

[83] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732, 2018.

[84] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1426–1435, 2019.