Visual Domain Adaptation with Manifold Embedded Distribution Alignment

Jindong Wang, Wenjie Feng, Yiqiang Chen†
Institute of Computing Technology, CAS, Beijing, China
{wangjindong,yqchen}@ict.ac.cn
fengwenjie@software.ict.ac.cn

Meiyu Huang
Qian Xuesen Lab. of Space Technology, China Academy of
Space Technology, Beijing, China
huangmeiyu@qxslab.cn

Han Yu‡
School of Computer Science and Engineering, NTU
Singapore
han.yu@ntu.edu.sg

Philip S. Yu§
Department of Computer Science, UIC
Chicago, USA
psyu@uic.edu

ABSTRACT
Visual domain adaptation aims to learn robust classifiers for the
target domain by leveraging knowledge from a source domain. Exist-
ing methods either attempt to align the cross-domain distributions,
or perform manifold subspace learning. However, there are two
significant challenges: (1) degenerated feature transformation, which
means that distribution alignment is often performed in the original
feature space, where feature distortions are hard to overcome. On
the other hand, subspace learning is not sufficient to reduce the dis-
tribution divergence. (2) unevaled distribution alignment, which
means that existing distribution alignment methods only align
the marginal and conditional distributions with equal importance,
while they fail to evaluate the different importance of these two
distributions in real applications. In this paper, we propose a Mani-
fold Embedded Distribution Alignment (MEDA) approach to
address these challenges. MEDA learns a domain-invariant clas-
sifier in Grassmann manifold with structural risk minimization,
while performing dynamic distribution alignment to quantitatively
account for the relative importance of marginal and conditional
distributions. To the best of our knowledge, MEDA is the first
attempt to perform dynamic distribution alignment for manifold domain
adaptation. Extensive experiments demonstrate that MEDA shows
significant improvements in classification accuracy compared to
state-of-the-art traditional and deep methods.

†The first two authors contributed equally.
‡Corr. to Y. Chen. J. Wang and Y. Chen are also affiliated with Beijing Key Lab. of Mobile
Computing and Pervasive Devices. W. Feng is also with CAS Key Lab. of Network
Data Science & Technology. J. Wang and W. Feng are also with University of Chinese
Academy of Sciences, Beijing, China.
§H. Yu is also with the Joint NTU-UBC Research Centre of Excellence in Active Living
for the Elderly (LILY), and the Alibaba-NTU Singapore Joint Research Institute.
P. Yu is also affiliated with Institute for Data Science, Tsinghua University, Beijing,
China.

Permission to make digital or hard copies of all or part of this work for personal or
classroom use is granted without fee provided that copies are not made or distributed
for profit or commercial advantage and that copies bear this notice and the full citation
on the first page. Copyrights for components of this work owned by others than ACM
must be honored. Abstracting with credit is permitted. To copy otherwise, or republish,
to post on servers or to redistribute to lists, requires prior specific permission and/or a
fee. Request permissions from permissions@acm.org.

MM ’18, Oct. 22–26, 2018, Seoul, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5665-7/18/10 .. $15.00
https://doi.org/10.1145/3240508.3240512

1 INTRODUCTION
The rapid growth of online media and content sharing applica-
tions has stimulated a great demand for automatic recognition
and analysis for images and other multimedia data [8, 20]. Un-
fortunately, it is often expensive and time-consuming to acquire
sufficient labeled data to train machine learning models. Thus, it is
often necessary to leverage the abundant labeled samples in some
existing domains to facilitate learning in a new target domain. Do-
main adaptation [27, 36] has been a promising approach to solve
such cross-domain learning problems.

Since the distributions of the source and target domains are
different, the key to successful adaptation is to reduce the distri-
bution divergence. To this end, existing work can be summarized
into two main categories: (a) instance reweighting [9, 39], which
reuses samples from the source domain according to some weight-
ing technique; and (b) feature matching, which either performs
subspace learning by exploiting the subspace geometrical struc-
ture [13, 15, 30], or distribution alignment to reduce the marginal or
conditional distribution divergence between domains [23, 40]. Our
focus is on feature matching methods. There are two significant
challenges in existing methods, i.e. degenerated feature transforma-
tion and unevaled distribution alignment.

Degenerated feature transformation means that both subspace
learning and distribution alignment can only reduce, but not remove
the distribution divergence [1]. Specifically, subspace learning [13,
15, 30] conducts subspace transformation to obtain better feature
representations. However, feature divergence is not eliminated after
subspace transformation [22] since subspace learning only utilizes
the subspace or manifold structure, but fails to perform feature
alignment. On the other hand, distribution alignment [23, 26, 37]
usually reduces the distribution distance in the original feature space, where features are often distorted [3] which makes it hard to reduce the divergence between domains. Therefore, it is critical to exploit both the advantages of subspace learning and distribution alignment to further facilitate domain adaptation.

Unevaluated distribution alignment means that existing work [19, 23, 33, 40] only attempted to align the marginal and conditional distributions with equal weights. But they failed to evaluate the relative importance of these two distributions. For example, when two domains are very dissimilar (Figure 1(a) → 1(b)), the marginal distribution is more important to align. When the marginal distributions are close (Figure 1(a) → 1(c)), the conditional distribution should be given more weight. However, there is no alignment method which can quantitatively account for the importance of these two distributions in conjunction.

As far as we know, there has been no previous work that tackle these two challenges together. In this paper, we propose a novel Manifold Embedded Distribution Alignment (MEDA) method to address the challenges of both degenerated feature transformation and unevaluated distribution alignment. MEDA learns a domain-invariant classifier in Grassmann manifold with structural risk minimization, while performing dynamic distribution alignment by considering the different importance of marginal and conditional distributions. We also provide a feasible solution to quantitatively evaluate the importance of distributions. To our best knowledge, MEDA is the first attempt to reveal the relative importance of marginal and conditional distributions in transfer learning.

This work makes the following contributions:

1) We propose the MEDA approach for domain adaptation. MEDA is capable of addressing both the challenges of degenerated feature transformation and unevaluated distribution alignment.

2) We propose the first quantitative evaluation of the relative importance of marginal and conditional distributions in domain adaptation. This is significantly useful in future research on transfer learning.

3) Extensive experiments on 7 real-world image datasets demonstrate that compared to several state-of-the-art traditional and deep methods, MEDA achieves a significant improvement of 3.5% in average classification accuracy.

2 RELATED WORK

MEDA substantially distinguishes from existing feature matching domain adaptation methods in several aspects:

Subspace learning. Subspace Alignment (SA) [13] aligned the base vectors of both domains, but failed to adapt feature distributions. Subspace distribution alignment (SDA) [31] extended SA by adding the subspace variance adaptation. However, SDA did not consider the local property of subspaces and ignored conditional distribution alignment. CORAL [30] aligned subspaces in second-order statistics, but it did not consider the distribution alignment. Scatter component analysis (SCA) [14] converted the samples into a set of subspaces (i.e. scatters) and then minimized the divergence between them. GFK [15] extended the idea of sampled points in manifold [16] and proposed to learn the geodesic flow kernel between domains. The work of [4] used a Hellinger distance to approximate the geodesic distance in Riemann space. [3] proposed to use Grassmann for domain adaptation, but they ignored the conditional distribution alignment. Different from these approaches, MEDA can learn a domain-invariant classifier in the manifold and align both marginal and conditional distributions.

Distribution alignment. MEDA substantially differs from existing work that only align marginal or conditional distribution [26]. Joint distribution adaptation (JDA) [23] proposed to match both distributions with equal weights. Others extended JDA by adding regularization [22], sparse representation [38], structural consistency [19], domain invariant clustering [33], and label propagation [40]. The main differences between MEDA and these methods are: 1) These work treats the two distributions equally. However, when there is a greater discrepancy between both distributions, they cannot evaluate their relative importance and thus lead to undermined performance. Our work is capable of evaluating the quantitative importance of each distribution via considering their different effects. 2) These methods are designed only in the original space, where feature distortion will hinder the performance. MEDA can align the distributions in the manifold to overcome the feature distortions.

Domain-invariant classifier learning. The recent work of ARTL [22], DIP [2, 3], and DMM [7] also aimed to build a domain-invariant classifier. However, ARTL and DMM can be undermined by feature distortion in original space, and they failed to leverage the different importance of distributions. DIP mainly focused on feature transformation and only aligned marginal distributions. MEDA is able to avoid the feature distortion and quantitatively evaluate the importance of marginal and conditional distribution alignment.

3 MANIFOLD EMBEDDED DISTRIBUTION ALIGNMENT

3.1 Problem Definition

Given a labeled source domain \(\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^n\) and an unlabeled target domain \(\mathcal{D}_t = \{(x'_j)_{j=m+1}^n\}\), assume the feature space \(X_s = X_t\), label space \(Y_s = Y_t\), but marginal probability \(P_s(x_s) \neq P_t(x_t)\) with conditional probability \(Q_s(y_s|x_s) \neq Q_t(y_t|x_t)\). The goal of domain adaptation is to learn a classifier \(f : x_s \rightarrow y_s\) to predict the labels \(y_t \in Y_t\) for the target domain \(\mathcal{D}_t\) using labeled source domain \(\mathcal{D}_s\).

According to the structural risk minimization (SRM) [35], \(f^* = \arg\min_{f \in \mathcal{H}_K} \|f(x,y) + R(f)\|_K\), where the first term indicates the loss on data samples, the second term denotes the regularization term, and \(\mathcal{H}_K\) is the Hilbert space induced by kernel function \(K(\cdot, \cdot)\). Since there is no labels on \(\mathcal{D}_t\), we can only perform SRM on \(\mathcal{D}_s\). Moreover, due to the different distributions between \(\mathcal{D}_s\) and \(\mathcal{D}_t\), it is necessary to add other constraints to maximize the distribution consistency while learning \(f\).
which we embed Geodesic Flow Kernel (GFK) [15] to learn the geodesic flow \( \Phi(t) : 0 \leq t \leq 1 \) between two points can draw a path for the two subspaces. If we let \( S_s = \Phi(0) \) and \( S_t = \Phi(1) \), then finding a geodesic flow from \( \Phi(0) \) to \( \Phi(1) \) equals to transforming the original features into an infinite-dimensional feature space, which eventually eliminates the domain shift. This kind of approach can be seen as an incremental way of ‘walking’ from \( \Phi(0) \) to \( \Phi(1) \). Specifically, the new features can be represented as \( z = g(x) = \Phi(t)^T x \). From [15], the inner product of transformed features \( z_i \) and \( z_j \) gives rise to a positive semidefinite geodesic flow kernel:

\[
\langle z_i, z_j \rangle = \int_0^1 (\Phi(t)^T x_i) (\Phi(t)^T x_j) \, dt = x_i^T G x_j \quad (2)
\]

Thus, the feature in original space can be transformed into Grassmann manifold with \( z = g(x) = \sqrt{G} x \). G can be computed efficiently by singular value decomposition [15]. Note that \( \sqrt{G} \) is only an expression form and cannot be computed directly, while its square root is calculated by Denman-Beavers algorithm [10].

### 3.4 Dynamic Distribution Alignment

The purpose of dynamic distribution alignment is to quantitatively evaluate the importance of aligning marginal (P) and conditional (Q) distributions in domain adaptation. Existing methods [23, 40] failed in this evaluation by only assuming that both distributions are equally important. However, this assumption may not be realistic for real applications. For instance, when transferring from Figure 1(a) to 1(b), there is a large difference between datasets. Therefore, the divergence between \( P_s \) and \( P_t \) is more dominant. In contrast, from Figure 1(a) to 1(c), the datasets are similar. Therefore, the distribution divergence in each class \( (Q_s \) and \( Q_t \) is more dominant.

**The adaptive factor:**

In view of this phenomenon, we introduce an adaptive factor to dynamically leverage the importance of these two distributions. Formally, the dynamic distribution alignment \( \overline{D_f} \) is defined as

\[
\overline{D_f}(D_s, D_t) = (1 - \mu)D_f(P_s, P_t) + \mu \sum_{c=1}^{C} D_f^{(c)}(Q_s, Q_t) \quad (3)
\]

where \( \mu \in [0, 1] \) is the adaptive factor and \( c \in \{1, \cdots, C\} \) is the class indicator. \( D_f(P_s, P_t) \) denotes the marginal distribution alignment, and \( D_f^{(c)}(Q_s, Q_t) \) denotes the conditional distribution alignment for class \( c \).

When \( \mu \to 0 \), it means that the distribution distance between the source and the target domains is large. Thus, marginal distribution alignment is more important (Figure 1(a) \( \rightarrow \) 1(b)). When \( \mu \to 1 \), it means that feature distribution between domains is relatively small, so the distribution of each class is dominant. Thus, the conditional distribution alignment is more important (Figure 1(a) \( \rightarrow \) 1(c)). When \( \mu = 0.5 \), both distributions are treated equally as in existing methods [23, 40]. Hence, the existing methods can be regarded as the special cases of MEDA. By learning the optimal adaptive factor \( \nu_{opt} \) (which we will discuss later), MEDA can be applied to different domain adaptation problems.
We use the **mean squared error (MMD)** [6] to empirically calculate the distribution divergence between domains. As a nonparametric measurement, MMD has been widely applied in many existing methods [14, 26, 40], and its theoretical effectiveness has been verified in [17]. The MMD distance between distributions \( p \) and \( q \) is defined as \( d^2(p, q) = \mathbb{E}_p[\phi(z)] - \mathbb{E}_q[\phi(z)] \) where \( H_K \) is the reproducing kernel Hilbert space (RKHS) induced by feature map \( \phi() \). Here, \( \mathbb{E}[\cdot] \) denotes the mean of the embedded samples. In order to compute an MMD associated with \( f \), we adopt the *projected MMD* [28] and compute the marginal distribution alignment as \( D_f(P_s, P_t) = \|\mathbb{E}[f(z_s)] - \mathbb{E}[f(z_t)]\|^2_{H_K} \). Similarly, the conditional distribution alignment is \( D_f^C(Q_s, Q_t) = \|\mathbb{E}[f(x_c)] - \mathbb{E}[f(x_c^t)]\|^2_{H_K} \). Then, dynamic distribution alignment can be expressed as

\[
D_f(D_s, D_t) = (1 - \mu)\|\mathbb{E}[f(z_s)] - \mathbb{E}[f(z_t)]\|^2_{H_K} + \mu \sum_{c=1}^C \|\mathbb{E}[f(x_c^s)] - \mathbb{E}[f(x_c^t)]\|^2_{H_K}
\]

(4)

Note that since \( D_t \) has no labels, it is not feasible to evaluate the conditional distribution \( Q_t = Q_t(y_t|x_t) \). Instead, we follow the idea in [37] and use the class conditional distribution \( Q_t(x_t|y_t) \) to approximate \( Q_t \). In order to evaluate \( Q_t(x_t|y_t) \), we apply prediction to \( D_t \) using a base classifier trained on \( D_s \) to obtain soft labels for \( D_t \). The soft labels may be less reliable, so we iteratively refine the prediction. Note that we only use the base classifier in the first iteration. After that, MEDA can automatically refine the labels for \( D_t \) using results from previous iteration.

### The quantitative evaluation of the adaptive factor \( \mu \):

We can treat \( \mu \) as a parameter and tune its value by cross-validation techniques. However, there is no label for the target domain in unsupervised domain adaptation problems. It is extremely hard to calculate the value of \( \mu \). In this work, we made the first attempt towards calculating \( \mu \) (i.e. \( \mu \)) by exploiting the global and local structure of domains. We adopted the \( \mathcal{A} \)-distance [6] as the basic measurement. The \( \mathcal{A} \)-distance is defined as the error of building a linear classifier to discriminate two domains (i.e. a binary classification). Formally, we denote \( e(h) \) the error of a linear classifier \( h \) discriminating the two domains \( D_s \) and \( D_t \). Then, the \( \mathcal{A} \)-distance can be defined as

\[
d_A(D_s, D_t) = 2(1 - e(h))
\]

(5)

We can directly compute the marginal \( \mathcal{A} \)-distance using above equation, which is denoted as \( d_M \). For the \( \mathcal{A} \)-distance between conditional distributions, we denote \( d_c \) the \( \mathcal{A} \)-distance for the \( c \)th class. It can be calculated as \( d_c = d_A(D_s^{(c)}, D_t^{(c)}) \), where \( D_s^{(c)} \) and \( D_t^{(c)} \) denote samples from class \( c \) in \( D_s \) and \( D_t \), respectively. Eventually, \( \mu \) can be estimated as

\[
\hat{\mu} = 1 - \frac{d_M}{d_M + \sum_{c=1}^C d_c}
\]

(6)

This estimation has to be conducted at every iteration of the dynamic distribution adaptation, since the feature distribution may vary after evaluating the conditional distribution each time. To be noticed, this is the first solution to quantitatively estimate the relative importance of each distribution. In fact, this estimation can be of significant help in future research on transfer learning and domain adaptation.

### 3.5 Learning Classifier \( f \)

After manifold feature learning and dynamic distribution alignment, \( f \) can be learned by summing up the SRM over \( D_s \) and distribution alignment. Adopting the square loss \( l_2 \), \( f \) can be represented as

\[
f = \arg \min_{f \in H_K} \sum_{i=1}^n (y_i - f(x_i))^2 + \eta \|f\|_K^2
\]

(7)

\[
+ \lambda D_f(D_s, D_t) + \rho R_f(D_s, D_t)
\]

In order to perform efficient learning, we now reformulate each term in detail.

#### SRM on the Source Domain:

Using the representer theorem [5], \( f \) admits the expansion

\[
f(z) = \sum_{i=1}^{n+m} \beta_i K(z_i, z)
\]

(8)

where \( \beta = (\beta_1, \beta_2, \cdots)^T \in \mathbb{R}^{(n+m)\times 1} \) is the coefficients vector and \( K \) is a kernel. Then, SRM on \( D_s \) can be

\[
\sum_{i=1}^n (y_i - f(z_i))^2 + \eta \|f\|_K^2
\]

(9)

\[
= \sum_{i=1}^{n+m} A_{ii}(y_i - f(z_i))^2 + \eta \|f\|_K^2
\]

\[
= ||(y - \hat{\beta}^TK\hat{A})\|_F^2 + \eta tr(\hat{\beta}^TK\hat{\beta})
\]

where \( || \cdot ||_F \) is the Frobenious norm, \( K \in \mathbb{R}^{(n+m)\times(n+m)} \) is the kernel matrix with \( K_{ij} = K(z_i, z_j) \), and \( \hat{A} \in \mathbb{R}^{(n+m)\times(n+m)} \) is a diagonal domain indicator matrix with \( A_{ii} = 1 \) if \( i \in D_s \), otherwise \( A_{ii} = 0 \). \( Y = \{y_1, \cdots, y_{n+m}\} \) is the label matrix from source and the target domains. \( tr(\cdot) \) denotes the trace operation. Although the labels for \( D_t \) are unavailable, they can be filtered out by the indicator matrix \( A \).

#### Dynamic distribution alignment:

Using the representer theorem and kernel tricks, dynamic distribution alignment in equation (4) becomes

\[
D_f(D_s, D_t) = tr(\hat{\beta}^TKMK\hat{\beta})
\]

(10)

where \( M = (1 - \mu)M_o + \mu \sum_{c=1}^C M_c \) is the MMD matrix with its element calculated by

\[
(M_o)_{ij} = \begin{cases} 
\frac{1}{n_s}, & z_i, z_j \in D_s \\
\frac{1}{m_t}, & z_i, z_j \in D_t \\
0, & \text{otherwise}
\end{cases}
\]

(11)

\[
(M_c)_{ij} = \begin{cases} 
\frac{1}{n_c}, & z_i, z_j \in D_s^{(c)} \\
\frac{1}{m_c}, & z_i, z_j \in D_t^{(c)} \\
0, & \text{otherwise}
\end{cases}
\]

(12)

where \( n_c = |D_s^{(c)}| \) and \( m_c = |D_t^{(c)}| \).
Algorithm 1 Manifold Embedded Distribution Alignment

**Input:** Data matrix $X = [X_1, X_2]$, source domain labels $y_s$, manifold subspace dimension $d$, regularization parameters $\lambda, \eta, \rho$, and #neighbor $p$.

**Output:** Classifier $f$.

1. Learn manifold feature transformation kernel $G$ via equation (2), and get manifold feature $Z = \sqrt{G}X$.
2. Train a base classifier using $D_s$, then apply prediction on $D_t$ to get its soft labels $\hat{y}_t$.
3. Construct kernel $K$ using transformed features $Z_s = Z_{1:n_s}$ and $Z_t = Z_{n_1:n + m}$.
4. repeat
5. Calculate the adaptive factor $\hat{\mu}$ using equation (6), and compute $M_0$ and $M_c$ by equations (11) and (12).
6. Compute $\hat{\beta}^*$ by solving equation (16) and obtain $f$ via the representer theorem in equation (8).
7. Update the soft labels of $D_t$: $\hat{y}_t = f(Z_t)$.
8. until Convergence
9. return Classifier $f$.

**Laplace Regularization:** Additionally, we add a Laplacian regularization term to further exploit the similar geometrical property of nearest points in manifold $D$ [5]. We denote the pair-wise affinity matrix as

$$W_{ij} = \begin{cases} \text{sim}(z_i, z_j), & z_i \in N_p(z_j) \text{ or } z_j \in N_p(z_i) \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (13)

where $\text{sim}(\cdot, \cdot)$ is a similarity function (such as cosine distance) to measure the distance between two points. $N_p(z_i)$ denotes the set of $p$-nearest neighbors to point $z_i$. $p$ is a free parameter and must be set in the method. By introducing Laplacian matrix $L = D - W$ with diagonal matrix $D_{ii} = \sum_{j=1}^{n+m} W_{ij}$, the final regularized can be expressed by

$$R_f(D_s, D_t) = \sum_{i,j=1}^{n+m} W_{ij}(f(z_i) - f(z_j))^2$$

$$= \sum_{i,j=1}^{n+m} f(z_i)L_{ij}f(z_j)$$

$$= \text{tr} \left( \hat{\beta}^T K L K \hat{\beta} \right)$$  \hspace{1cm} (14)

**Overall Reformulation:** Substituting with equations (9), (10) and (14), $f$ in equation (7) can be reformulated as

$$f = \arg \min_{f \in H_K} \| (Y - \beta^T K)A \|_F^2 + \eta \text{tr}(\beta^T K \beta)$$

$$+ \text{tr} \left( \beta^T (\lambda M + \rho L) \beta \right)$$  \hspace{1cm} (15)

Setting derivative $\partial f / \partial \beta = 0$, we obtain the solution

$$\hat{\beta}^* = ((A + \lambda M + \rho L)K + \eta I)^{-1}AY^T$$  \hspace{1cm} (16)

MEDA has a nice property: it can learn the cross-domain function directly without the need of explicit classifier training. This makes it significantly different from most existing work such as JGSA [40] and CORAL [30] that further needs to learn a certain classifier.

4 EXPERIMENTS AND EVALUATIONS

In this section, we evaluate the performance of MEDA through extensive experiments on large-scale public datasets. The source code for MEDA is available at http://transferlearning.xyz/.

4.1 Data Preparation

We adopted seven publicly available datasets: Office+Caltech10, USPS + MNIST, ImageNet + VOC2007, and Office-31. These datasets are popular for benchmarking domain adaptation algorithms and have been widely adopted in most existing work such as [15, 22, 40, 41]. Table 1 lists the statistics of the seven datasets.

**Office-31** [29] consists of three real-world object domains: Amazon (A), Webcam (W) and DSLR (D). It has 4,652 images with 31 categories. **Caltech-256** (C) contains 30,607 images and 256 categories. Since the objects in Office and Caltech follow different distributions, domain adaptation can help to perform cross-domain recognition. There are 10 common classes in the two datasets. For our experiments, we adopted the **Office-Caltech10** datasets from [15] which contains 12 tasks: $A \rightarrow D$, $A \rightarrow C$, ..., $C \rightarrow W$. In the rest of the paper, we use $A \rightarrow B$ to denote the knowledge transfer from source domain $A$ to the target domain $B$.

**USPS** (U) and **MNIST** (M) are standard digit recognition datasets containing handwritten digits from 0-9. Since the same digits across two datasets follow different distributions, it is necessary to perform domain adaptation. USPS consists of 7,929 training images and 2,097 test images of size $16 \times 16$. MNIST consists of 60,000 training images and 10,000 test images of size $28 \times 28$. We construct two tasks: $U \rightarrow M$ and $M \rightarrow U$.

**ImageNet** (I) and **VOC2007** (V) are large standard image recognition datasets. Each dataset can be treated as one domain. The images from the same classes of two domains follow different distributions. In our experiments, we adopt the sub-datasets presented in [12] to construct cross-domain tasks. Five common classes are extracted from both datasets: bird, cat, chair, dog, and person. Eventually, we have two tasks: $I \rightarrow V$ and $V \rightarrow I$.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Sample</th>
<th>#Feature</th>
<th>#Class</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office-10</td>
<td>1,410</td>
<td>800 (4,096)</td>
<td>10</td>
<td>A, W, D</td>
</tr>
<tr>
<td>Caltech-10</td>
<td>1,123</td>
<td>800 (4,096)</td>
<td>10</td>
<td>C</td>
</tr>
<tr>
<td>Office-31</td>
<td>4,652</td>
<td>4,096</td>
<td>31</td>
<td>A, W, D</td>
</tr>
<tr>
<td>USPS</td>
<td>1,800</td>
<td>256</td>
<td>10</td>
<td>USPS (U)</td>
</tr>
<tr>
<td>MNIST</td>
<td>2,000</td>
<td>256</td>
<td>10</td>
<td>MNIST (M)</td>
</tr>
<tr>
<td>ImageNet</td>
<td>7,341</td>
<td>4,096</td>
<td>5</td>
<td>ImageNet (I)</td>
</tr>
<tr>
<td>VOC2007</td>
<td>3,376</td>
<td>4,096</td>
<td>5</td>
<td>VOC (V)</td>
</tr>
</tbody>
</table>

4.2 State-of-the-art Comparison Methods

We compared the performance of MEDA with several state-of-the-art traditional and deep domain adaptation approaches.

**Traditional learning methods:**
- 1NN, SVM, and PCA
- Transfer Component Analysis (TCA) [26], which performs marginal distribution alignment
- Geodesic Flow Kernel (GFK) [15], which performs manifold feature learning
- Joint distribution alignment (JDA) [23], which adapts both marginal and conditional distribution
<table>
<thead>
<tr>
<th>Task</th>
<th>INN</th>
<th>SVM</th>
<th>PCA</th>
<th>TCA</th>
<th>GFK</th>
<th>JDA</th>
<th>TJM</th>
<th>CORAL</th>
<th>SCA</th>
<th>ARTL</th>
<th>JGSA</th>
<th>MEDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>C → A</td>
<td>23.7</td>
<td>33.1</td>
<td>39.5</td>
<td>45.6</td>
<td>46.0</td>
<td>43.1</td>
<td>46.8</td>
<td>52.1</td>
<td>45.6</td>
<td>44.1</td>
<td>51.5</td>
<td>56.5</td>
</tr>
<tr>
<td>C → W</td>
<td>25.8</td>
<td>41.7</td>
<td>34.6</td>
<td>39.3</td>
<td>37.0</td>
<td>39.5</td>
<td>39.0</td>
<td>46.4</td>
<td>40.0</td>
<td>31.5</td>
<td>45.4</td>
<td>53.9</td>
</tr>
<tr>
<td>C → D</td>
<td>25.3</td>
<td>47.8</td>
<td>44.6</td>
<td>45.9</td>
<td>40.8</td>
<td>49.0</td>
<td>44.6</td>
<td>45.9</td>
<td>47.1</td>
<td>39.5</td>
<td>45.9</td>
<td>50.3</td>
</tr>
<tr>
<td>A → C</td>
<td>26.0</td>
<td>41.7</td>
<td>39.0</td>
<td>42.0</td>
<td>40.7</td>
<td>40.9</td>
<td>39.3</td>
<td>45.1</td>
<td>39.7</td>
<td>36.1</td>
<td>41.5</td>
<td>43.9</td>
</tr>
<tr>
<td>A → W</td>
<td>29.8</td>
<td>31.9</td>
<td>35.9</td>
<td>40.0</td>
<td>37.0</td>
<td>38.0</td>
<td>42.0</td>
<td>44.4</td>
<td>34.9</td>
<td>33.6</td>
<td>45.8</td>
<td>53.2</td>
</tr>
<tr>
<td>A → D</td>
<td>25.3</td>
<td>44.6</td>
<td>33.8</td>
<td>35.7</td>
<td>40.1</td>
<td>42.0</td>
<td>45.2</td>
<td>39.5</td>
<td>39.5</td>
<td>36.9</td>
<td>47.1</td>
<td>45.9</td>
</tr>
<tr>
<td>W → C</td>
<td>19.9</td>
<td>28.8</td>
<td>28.2</td>
<td>31.5</td>
<td>24.8</td>
<td>33.0</td>
<td>30.2</td>
<td>32.3</td>
<td>31.1</td>
<td>29.7</td>
<td>33.2</td>
<td>34.0</td>
</tr>
<tr>
<td>W → A</td>
<td>23.0</td>
<td>27.6</td>
<td>29.1</td>
<td>30.5</td>
<td>27.6</td>
<td>29.8</td>
<td>30.0</td>
<td>36.0</td>
<td>30.0</td>
<td>38.3</td>
<td>39.9</td>
<td>42.7</td>
</tr>
<tr>
<td>W → D</td>
<td>39.2</td>
<td>78.3</td>
<td>89.2</td>
<td>91.1</td>
<td>85.4</td>
<td>92.4</td>
<td>89.2</td>
<td>86.6</td>
<td>87.3</td>
<td>87.9</td>
<td>90.5</td>
<td>88.5</td>
</tr>
<tr>
<td>D → C</td>
<td>26.3</td>
<td>26.4</td>
<td>29.7</td>
<td>33.0</td>
<td>29.3</td>
<td>31.2</td>
<td>31.4</td>
<td>33.8</td>
<td>30.7</td>
<td>30.5</td>
<td>29.9</td>
<td>34.9</td>
</tr>
<tr>
<td>D → A</td>
<td>28.5</td>
<td>26.2</td>
<td>33.2</td>
<td>32.8</td>
<td>28.7</td>
<td>33.4</td>
<td>32.8</td>
<td>37.7</td>
<td>31.6</td>
<td>34.9</td>
<td>38.0</td>
<td>41.2</td>
</tr>
<tr>
<td>D → W</td>
<td>63.4</td>
<td>52.5</td>
<td>86.1</td>
<td>87.5</td>
<td>80.3</td>
<td>89.2</td>
<td>85.4</td>
<td>84.7</td>
<td>84.4</td>
<td>88.5</td>
<td>91.9</td>
<td>87.5</td>
</tr>
<tr>
<td>Average</td>
<td>31.4</td>
<td>41.1</td>
<td>43.6</td>
<td>46.2</td>
<td>43.1</td>
<td>46.8</td>
<td>46.3</td>
<td>48.8</td>
<td>45.2</td>
<td>44.3</td>
<td>50.0</td>
<td>52.7</td>
</tr>
</tbody>
</table>

Table 3: Accuracy (%) on USPS+MNIST and ImageNet+VOC2007 datasets.

<table>
<thead>
<tr>
<th>Task</th>
<th>INN</th>
<th>SVM</th>
<th>PCA</th>
<th>TCA</th>
<th>GFK</th>
<th>JDA</th>
<th>TJM</th>
<th>CORAL</th>
<th>SCA</th>
<th>ARTL</th>
<th>JGSA</th>
<th>MEDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>U → M</td>
<td>44.7</td>
<td>62.2</td>
<td>45.0</td>
<td>51.2</td>
<td>46.5</td>
<td>59.7</td>
<td>52.3</td>
<td>50.5</td>
<td>48.0</td>
<td>67.7</td>
<td>68.2</td>
<td>72.1</td>
</tr>
<tr>
<td>M → U</td>
<td>65.9</td>
<td>68.2</td>
<td>66.2</td>
<td>66.3</td>
<td>61.2</td>
<td>67.5</td>
<td>63.3</td>
<td>49.2</td>
<td>65.1</td>
<td>88.8</td>
<td>80.4</td>
<td>89.9</td>
</tr>
<tr>
<td>I → U</td>
<td>50.8</td>
<td>63.4</td>
<td>58.4</td>
<td>63.7</td>
<td>59.5</td>
<td>63.4</td>
<td>63.3</td>
<td>59.6</td>
<td>-</td>
<td>62.4</td>
<td>52.3</td>
<td>67.8</td>
</tr>
<tr>
<td>V → I</td>
<td>38.2</td>
<td>42.7</td>
<td>65.1</td>
<td>64.9</td>
<td>73.8</td>
<td>70.2</td>
<td>73.0</td>
<td>70.3</td>
<td>-</td>
<td>72.2</td>
<td>70.6</td>
<td>74.7</td>
</tr>
<tr>
<td>Average</td>
<td>49.9</td>
<td>56.3</td>
<td>58.7</td>
<td>59.0</td>
<td>60.2</td>
<td>65.1</td>
<td>63.1</td>
<td>52.4</td>
<td>-</td>
<td>72.8</td>
<td>67.9</td>
<td>75.9</td>
</tr>
</tbody>
</table>

Table 4: Accuracy (%) on Office+Caltech10 datasets using DeCaf6 features.

- Transfer Joint Matching (TJM) [24], which adapts marginal distribution with source sample selection
- Adaptation Regularization (ARTL) [22], which learns domain classifier in original space
- CORelation Alignment (CORAL) [30], which performs second-order subspace alignment
- Scatter Component Analysis (SCA) [14], which adapts scatterers in subspace
- Joint Geometrical and Statistical Alignment (JGSA) [40], which aligns marginal & conditional distributions with label propagation
- Distribution Matching Machine (DMM) [7], which learns a transfer SVM to align distributions

And deep domain adaptation methods:
- AlexNet [21], which is a standard convnet
- Deep Domain Confusion (DDC) [34], which is a single-layer deep adaptation method with MMD loss
- Deep Adaptation Network (DAN) [25], which is a multi-layer adaptation method with multiple kernel MMD
- Deep CORAL (DCORAL) [32], which is a deep neural network with CORAL loss
- Deep Unsupervised Convolutional Domain Adaptation (DUCDA) [41], which is based on attention and CORAL loss

4.3 Experimental Setup

For fair comparison, we follow the same protocols as [37, 40, 41] to adopt the extracted features for MEDA and other traditional methods. To be specific, 256 SURF features are used for USPS+MNIST datasets; for Office+Caltech10 datasets, both 800 SURF and 4,096 DeCaf6 features are used; for Office-31 dataset, 4,096 DeCaf6 features are used; for ImageNet+VOC datasets, 4,096 DeCaf6 features are used. Deep methods can be used to the original images.

Parameter setting: The optimal parameters of all comparison methods are set according to their original papers. As for MEDA, we set the manifold feature dimension $d = 20, 30, 40$ for Office+Caltech10, USPS+MNIST, and ImageNet+VOC datasets, respectively. The iteration number is set to $T = 10$. We use the RBF kernel with the bandwidth set to be the variance of inputs. The regularization parameters are set as $\rho = 10, \lambda = 10, \eta = 0.1$, and $\rho = 1$. The
We can make several observations as follows. Additionally, the experiments on parameter sensitivity and convergence analysis in later experiments (Section 4.6 and 4.7) indicate that MEDA stays robust with a wide range of parameter choices.

We adopt classification Accuracy as the evaluation metric, which is widely used in existing literatures [15, 26, 37]:

\[
\text{Accuracy} = \frac{\sum_{(x,y) \in D} \mathbb{1}(y(x) = y(\hat{x}))}{\sum_{x \in D} \mathbb{1}(x \in D)},
\]

where \(y(x)\) and \(\hat{y}(x)\) are the truth and predicted labels for target domain, respectively.

### 4.4 Experimental Results and Analysis

The classification accuracy results on the aforementioned datasets are shown in Tables 2, 3, and 4, respectively[^1]. From those results, we can make several observations as follows. Firstly, MEDA outperformed all other traditional and deep comparison methods in most tasks (21/28 tasks). The average classification accuracy of MEDA on 28 tasks was 73.2%. Compared to the best baseline method JGSA (69.7%), the average performance improvement was 3.5%, which showed a significant average error reduction of 11.6%. Note that the results on Office-31 dataset were in the supplementary file[^2] due to space constraints, and the observations are the same. Since these results were obtained from a wide range of image datasets, it demonstrates that MEDA is capable of significantly reducing the distribution divergence in domain adaptation problems.

Secondly, the performances of distribution alignment methods (TCA, JDA, ARTL, TJM, JGSA, and DMM) and subspace learning methods (GFK, CORAL, and SCA) were generally worse than MEDA. Each kind of methods has its limitations and cannot handle domain adaptation in specific tasks. This indicates the disadvantages of those methods to cope with degenerated feature transformation and unevaluated distribution alignment. After manifold or subspace learning, there still exists large domain shift [3]; while feature distortion will undermine the distribution alignment methods.

Thirdly, MEDA also outperformed the deep methods (AlexNet, DDC, DAN, DCORAL, and DUCDA) on Office+Caltech10 datasets. Deep methods often have to tune a lot of hyperparameters before obtaining the optimal results. Compared to them, MEDA only involves several parameters that can easily be set by human experience or cross-validation. This implies the accuracy and efficiency of MEDA in domain adaptation problems over other deep methods.

### 4.5 Effectiveness Analysis

#### 4.5.1 Manifold Feature Learning

We investigate the effectiveness of manifold feature learning in handling the degenerated feature transformation challenge. To this end, we ran MEDA with and without manifold feature learning on randomly selected tasks. Table 5 showed the mean, standard deviation, and performance improvement of classification accuracy with \(\mu \in \{0, 0.1, \cdots, 1\}\). For instance, the improvement of mean accuracy on task C \(\rightarrow\) A was: \((56.5 - 44.9)/44.9 \times 100\% = 25.8\%\). From these results, we can observe that: 1) The performance of all the tasks were improved with manifold feature learning, indicating that transforming features into the manifold alleviates domain shift to some extent and facilitates distribution alignment; 2) The standard deviation of methods that adopted manifold learning with different \(\mu\) could be dramatically reduced. 3) MEDA can also reach a comparable performance without manifold learning, while adding manifold learning would produce better results. This reveals the effectiveness of manifold feature learning to alleviate degenerated feature transformation.

#### 4.5.2 Dynamic Distribution Alignment

We verify the effectiveness of dynamic distribution alignment in handling the unevaluated distribution alignment challenge. We ran MEDA by searching \(\mu \in \{0, 0.1, \cdots, 0.9, 1.0\}\) and compared the performances with the best baseline method (JGSA). From the results in Figure 3, we can clearly observe that the classification accuracy varied with different choice of \(\mu\). This indicates the necessity to consider the different effects between marginal and conditional distributions. We can also observe that the optimal \(\mu\) value varied on different tasks (\(\mu = 0.2, 0.1\) for three tasks, respectively). Thus, it is necessary to dynamically adjust the distribution alignment between domains according to different tasks. Moreover, the optimal value of \(\mu\) is not unique on certain task. The classification results may be the same even for different \(\mu\).

---

[^1]: Symbol 'x' denotes the result is not available since there is no code or results.
[^2]: Supplementary file is at [https://www.jiangyuym.com/p/DRuWOFkQkJazdBRk2E](https://www.jiangyuym.com/p/DRuWOFkQkJazdBRk2E).
The estimation of $\mu$: We evaluate our solution of estimating $\mu$ (equation (6)). Since the optimal $\mu$ is not unique, we can not directly compare the value of $\mu_{opt}$ and $\tilde{\mu}$ to evaluate our solution. Instead, we compare the performances (accuracy values) achieved by $\mu_{opt}$ and $\tilde{\mu}$. The results in Table 6 indicated that the performance of estimated $\tilde{\mu}$ was very close to $\mu_{opt}$, and sometimes it is better than grid search ($M \rightarrow U$). For instance, the performance variation of $C \rightarrow A$ was $(57.0 - 56.5)/57.0 \times 100 \% = 0.9\%$. This demonstrates that the effectiveness in estimating $\mu$. This estimation solution can be directly applied to future research.

4.5.3 Evaluation of Each Component. When learning the final classifier $f$, MEDA involves three components: the structural risk minimization (SRM), the dynamic distribution alignment (DA), and Laplacian regularization (Lap). We empirically evaluated the importance of each component. We randomly selected several tasks and reported the results in Figure 4. Note that we did not run this experiment on the Decaf features of Office+Caltech10 dataset since its results are already satisfied.

Those results clearly indicated that each component is important in MEDA, and they are indispensable. Moreover, we observe that in all tasks, it is more important to align the distributions. The reason is that there exists large distribution divergence between two domains. The results also suggests that adding Laplacian regularization is more beneficial in capturing the manifold structure. Additionally, combining the effectiveness of manifold feature learning (Section 4.5.1), it is clear that all components are important for improving the accuracy in domain adaptation tasks.

4.6 Parameter Sensitivity

As with other state-of-the-art domain adaptation algorithms [14, 22, 40], MEDA also involves several parameters. In this section, we evaluate the parameter sensitivity.

4.6.1 Subspace Dimension and #neighbor. We investigated the sensitivity of manifold subspace dimension $d$ and #neighbor $p$ through experiments with a wide range of $d \in \{10, 20, \ldots, 100\}$ and $p \in \{2, 4, \ldots, 64\}$ on randomly selected tasks. From the results in Figure 5(a) and 5(b), it can be observed that MEDA was robust with regard to different values of $d$ and $p$. Therefore, they can be selected without knowledge in real applications.

4.6.2 Regularization Parameters. We ran MEDA with a wide range of values for regularization parameters $\lambda$, $\eta$, and $\rho$ on several random tasks and compare its performance with the best baseline method. For the lack of space, we only report the results of $\lambda$ in (c) and the results of $\rho$ and $\eta$ can be found in the supplementary file. We observed that MEDA can achieve a robust performance with regard to a wide range of parameter values. Specifically, the best choices of these parameters are: $\lambda \in \{0.5, 1.000\}$, $\eta \in \{0.01, 1\}$, and $\rho \in \{0.01, 5\}$. To sum up, the performance of MEDA stays robust with a wide range of regularization parameter choice.

4.7 Convergence and Time Complexity

We validated the convergence of MEDA through empirical analysis. From the results in Figure 5(d), it can be observed that MEDA can reach a steady performance in only a few ($T < 10$) iterations. It indicates the training advantage of MEDA in cross-domain tasks.

We also empirically checked the time complexity of MEDA and compared it with other top two baselines ARTL and JGSA on different tasks. The environment was an Intel Core i7-4790 CPU with 24 GB memory. Note that the time complexity of deep methods are not comparable with MEDA since they require a lot of backpropagations. The results in Table 7 reveal that except its superiority in classification accuracy, MEDA also achieved a running time complexity comparable to top two best baseline methods.

5 CONCLUSIONS

In this paper, we propose a novel Manifold Embedded Distribution Alignment (MEDA) approach for visual domain adaptation. Compared to existing work, MEDA is the first attempt to handle the challenges of both degenerated feature transformation and unevaluated distribution alignment. MEDA can learn the domain-invariant classifier with the principle of structural risk minimization while performing dynamic distribution alignment. We also provide a feasible solution to quantitatively calculate the adaptive factor. We conducted extensive experiments on several large-scale publicly available image classification datasets. The results demonstrate the superiority of MEDA against other state-of-the-art traditional and deep domain adaptation methods.

ACKNOWLEDGMENTS

This work is supported by National Key R & D Program of China (2016YFB1001200), NSFC (61572471,61702520,61672313), NSF through grants IIS-1526499, IIS-1763325, CNS-1626432, and Nanyang Assistant Professorship (NAP) of Nanyang Technological University.