Semi-Supervised Optimal Transport for Heterogeneous Domain Adaptation

Yuguang Yan[1*], **Wen Li**[2*], **Hanrui Wu**[1], **Huaqing Min**[1], **Mingkui Tan**[1†] and **Qingyao Wu**[1†]
[1]School of Software Engineering, South China University of Technology, China
[2]Computer Vision Laboratory, ETH Zurich, Switzerland
{yan.yuguang,sehrwu}@mail.scut.edu.cn, {mingkuitan,qyw,hqmin}@scut.edu.cn
liwen@vision.ee.ethz.ch

Abstract

Heterogeneous domain adaptation (HDA) aims to exploit knowledge from a heterogeneous source domain to improve the learning performance in a target domain. Since the feature spaces of the source and target domains are different, the transferring of knowledge is extremely difficult. In this paper, we propose a novel semi-supervised algorithm for HDA by exploiting the theory of optimal transport (OT), a powerful tool originally designed for aligning two different distributions. To match the samples between heterogeneous domains, we propose to preserve the semantic consistency between heterogeneous domains by incorporating label information into the entropic Gromov-Wasserstein discrepancy, which is a metric in OT for different metric spaces, resulting in a new semi-supervised scheme. Via the new scheme, the target and transported source samples with the same label are enforced to follow similar distributions. Lastly, based on the Kullback-Leibler metric, we develop an efficient algorithm to optimize the resultant problem. Comprehensive experiments on both synthetic and real-world datasets demonstrate the effectiveness of our proposed method.

1 Introduction

Domain adaptation (DA) aims to leverage data from an auxiliary source domain to assist the learning task in a domain of interest, *a.k.a.*, target domain. Most existing works of DA consider the homogeneous DA problem, in which source and target samples share the same feature space [Long *et al.*, 2014]. Compared with the homogeneous setting, heterogeneous domain adaptation (HDA) addresses a more challenging situation where the source and target samples are represented in different feature spaces [Li *et al.*, 2014]. For instance, source and target samples could be images embedded in different kinds of features [Yan *et al.*, 2017b], or documents in different languages [Zhou *et al.*, 2016].

Most of the previous studies of HDA are devoted to learning a feature mapping function to transform the source and target samples into a common feature space. Nevertheless, these mapping functions are usually restricted in a predefined assumption space, *e.g.*, linear projections used in [Tsai *et al.*, 2016; Yan *et al.*, 2017b], hence their capacities of transformation between two heterogeneous feature spaces would be rather limited.

In this paper, we propose a novel HDA algorithm by exploiting the theory of optimal transport (OT). Instead of learning a feature transformation, OT aims to match two distributions by transporting samples from one distribution to another. Very recently, a few works have been reported to apply OT for homogeneous domain adaptation [Perrot *et al.*, 2016; Courty *et al.*, 2017]. However, these methods rely on a transport distance metric for two sets of homogeneous samples, which is not applicable to HDA problems.

To address this issue, we treat the two heterogeneous feature spaces as two individual metric spaces and employ the entropic Gromov-Wasserstein (EGW) discrepancy [Peyré *et al.*, 2016] to learn a transport plan, which transports samples from one metric space into another metric space. The EGW discrepancy is a powerful metric in OT for learning an optimal transport matrix, such that the distance between two metric spaces is minimized. EGW has been shown to be effective for many tasks such as 3D object matching [Peyré *et al.*, 2016]. Since no common distance metric between two metric spaces is required, EGW can be naturally applied to the HDA problem.

A potential issue when applying EGW for HDA is that EGW is an unsupervised approach without considering label information. Therefore, even though the marginal distributions of two domains are matched, the semantic concepts may not be well aligned after the transportation. As illustrated in the simulated results in Figure 1, the transported source samples obtained by EGW (Figure 1(d)) follow a similar distribution with the target samples (Figure 1(b)). However, the target and transported source samples with different labels are mixed up, which leads to a significant performance drop in classification.

To alleviate the above issue, we propose to incorporate label information into EGW to learn an optimal transport plan

[*]The co-first author.
[†]The corresponding author.

satisfying label consistency for classification. We leverage both labeled and unlabeled target samples to constrain the transport matrix, such that the transported source samples and target samples with the same label follow similar distributions, as illustrated in Figure 1(f). The principal contributions of our work are summarized as follows.

- We propose a semi-supervised entropic Gromov-Wasserstein discrepancy approach named SGW to incorporate the supervision information when learning the optimal transport. A conditional distribution matching regularization and a group entropic regularization are introduced to effectively exploit the label information to constrain the transportation.
- We apply a projected gradient algorithm with the Kullback-Leibler divergence to efficiently solve the derived optimization problem.
- We conduct comprehensive experiments on both synthetic and real-world datasets to demonstrate the effectiveness of our proposed method.

2 Related Studies

By leveraging knowledge extracted from data in an auxiliary source domain, domain adaptation addresses the problem where the labeled samples in a target domain are insufficient to train an effective classifier [Pan and Yang, 2010; Patel *et al.*, 2015; Shao *et al.*, 2015]. Homogeneous domain adaptation considers the situation where the source and target domains share the same feature space but have different data distributions [Pan *et al.*, 2011; Long *et al.*, 2015; Yan *et al.*, 2017a; Liu *et al.*, 2017]. Recent years, heterogeneous domain adaptation (HDA) has attracted a lot of attention [Zhou *et al.*, 2014; 2016; Moon and Carbonell, 2017; Luo *et al.*, 2017]. Compared with the homogeneous setting, HDA is more challenging due to the heterogeneity between the source and target feature spaces, making it difficult to directly leverage source samples to assist the target learning task.

In general, existing HDA algorithms can be categorized into two groups. The first group uses extra auxiliary data to connect the source and target feature spaces. In [Wu *et al.*, 2014], co-occurrence data are employed to build relationships between the source and target samples, and graph-based algorithms are conducted for classification. TTL [Tan *et al.*, 2015] applies a collective matrix factorization method to learn semantic new representations for target data. Yan *et al.* exploited co-occurrence data to address HDA in online learning settings [Yan *et al.*, 2017c]. These methods require extra co-occurrence data, which are not always available in real-world applications.

Rather than relying on co-occurrence data to connect feature spaces, the second group of HDA algorithms aims to find effective feature transformations to map source and target samples into the same feature space. HFA [Duan *et al.*, 2012] and SHFA [Li *et al.*, 2014] augment source and target samples based on two projection matrices, and simultaneously train an SVM classifier on the augmented data. CDLS [Tsai *et al.*, 2016] finds representative landmarks to learn a domain-invariant feature subspace, and then a classifier for target data is trained in the learned subspace. DCA [Yan *et al.*, 2017b] jointly seeks for a discriminative correlation subspace defined by Canonical Correlation Analysis and learns a classifier in the found subspace.

Different from the above works, we address HDA by exploiting optimal transport [Peyré and Cuturi, 2017], which does not learn explicit feature transformations. Although there are a few works addressing homogeneous domain adaptation problems using optimal transport [Courty *et al.*, 2017; Perrot *et al.*, 2016; Redko *et al.*, 2017], optimal transport for HDA is challenging and barely studied in the literature. It is not applicable to directly use the existing transport cost for homogeneous samples to the HDA problem. In this paper, we propose to learn optimal transport for HDA by leveraging the entropic Gromov-Wasserstein discrepancy [Peyré *et al.*, 2016] to minimize the difference between the metric matrices of two heterogeneous domains.

3 Learning Model

3.1 Problem Statement and Notations

In heterogeneous domain adaptation (HDA), we are given a source domain and a target domain, which have different feature representations. The source domain contains a large number of labeled samples, and the target domain only contains a limited number of labeled samples (and a number of unlabeled samples in certain scenarios). The task is to take the advantage of the source labeled samples to improve the classification performance in the target domain.

For convenience of presentation, we denote the source samples as $\mathbf{X}_s = [\mathbf{x}_1^s, \ldots, \mathbf{x}_{n_s}^s]^\top \in \mathbb{R}^{n_s \times d_s}$ with n_s being the number of source samples and d_s being the source feature dimension, and $\mathbf{x}_i^s \in \mathbb{R}^{d_s}$ is the i-th source sample with label $y_i^s \in \mathcal{Y}_s$. The target samples are represented as $\mathbf{X}_t = [\mathbf{x}_1^t, \ldots, \mathbf{x}_{n_t}^t]^\top \in \mathbb{R}^{n_t \times d_t}$, where n_t is the number of target samples, $\mathbf{x}_i^t \in \mathbb{R}^{d_t}$ is the i-th target sample with d_t being the feature dimension, Among the target samples, $\mathbf{X}_l = [\mathbf{x}_1^l, \ldots, \mathbf{x}_{n_l}^l]^\top \in \mathbb{R}^{n_l \times d_t}$ are labeled target samples with labels $\{y_i^l\}_{i=1}^{n_l} \in \mathcal{Y}_t$, and $\mathbf{X}_u = [\mathbf{x}_1^u, \ldots, \mathbf{x}_{n_u}^u]^\top \in \mathbb{R}^{n_u \times d_t}$ are unlabeled target samples, where n_l and n_u are the number of labeled and unlabeled target samples, respectively, and $n_t = n_l + n_u$. In our HDA problems, We have $n_l \ll n_u$, $d_s \neq d_t$ and $\mathcal{Y}_s = \mathcal{Y}_t = \{1, \ldots, K\}$.

For the matrix $\mathbf{A}$, $\log(\mathbf{A})$ and $\exp(\mathbf{A})$ are element-wise operations. The entropy of $\mathbf{A}$ is defined by

$$H(\mathbf{A}) \stackrel{\text{def.}}{=} -\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} A_{i,j}(\log A_{i,j} - 1), \tag{1}$$

where $A_{i,j}$ is the (i,j)-th element of $\mathbf{A}$. For the matrices $\mathbf{A}$ and $\mathbf{B}$, the inner product is defined as

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_i \sum_j A_{i,j} B_{i,j}. \tag{2}$$

3.2 Entropic Gromov-Wasserstein Discrepancy

The major challenge in HDA is that the feature representations in two domains are different. Thus, it is important

to build the correspondence between heterogeneous feature spaces for reducing the domain discrepancy. To achieve this, we employ the entropic Gromov-Wasserstein (EGW) discrepancy [Peyré *et al.*, 2016], which has shown superior performance in 3D object matching.

The EGW discrepancy is based on the optimal transport theory, which seeks for an optimal solution to transport a set of samples (or a distribution) to another set of samples (or another distribution). The advantage of the EGW discrepancy is that it does not rely on the distance between two sets of samples, which is usually required in classical optimal transport problems. Instead, EGW is defined on two individual metric-measure spaces, thus is appropriate for the HDA problem, where the feature spaces of two domains are heterogeneous.

In particular, given a set of samples in one domain, the corresponding empirical distribution can be represented by a simplex of histograms $\mathbf{p} \in \Delta_n \overset{\text{def.}}{=} \{\mathbf{p} \in \mathbb{R}_+^n | \mathbf{p}^\top \mathbf{1}_n = 1\}$, where n is the number of samples, and the vector $\mathbf{1}_n = [1, \ldots, 1]^\top \in \mathbb{R}^n$. We use $\mathbf{p}_s = [\frac{1}{n_s}, \ldots, \frac{1}{n_s}]^\top$ and $\mathbf{p}_t = [\frac{1}{n_t}, \ldots, \frac{1}{n_t}]^\top$ to represent the empirical distributions of source and target samples, respectively. The joint distribution of $\mathbf{p}_s$ and $\mathbf{p}_t$ is defined by $\mathbf{T} \in \mathcal{T} \overset{\text{def.}}{=} \{\mathbf{T} \in \mathbb{R}^{n_s \times n_t} | \mathbf{T}\mathbf{1}_{n_t} = \mathbf{p}_s, \mathbf{T}^\top \mathbf{1}_{n_s} = \mathbf{p}_t\}$.

Let $\mathbf{M}_s$ (*resp.* $\mathbf{M}_t$) be a matrix representing a certain metric on source (*resp.* target) samples, the entropic Gromov-Wasserstein problem seeks for an optimal transport matrix $\mathbf{T}$, which gives the best match of two metric matrices. The EGW problem reads

$$\text{EGW}(\mathbf{M}_s, \mathbf{M}_t, \mathbf{p}_s, \mathbf{p}_t) \overset{\text{def.}}{=} \min_{\mathbf{T} \in \mathcal{T}} \mathcal{E}_{\mathbf{M}_s, \mathbf{M}_t}(\mathbf{T}) - \epsilon H(\mathbf{T}), \quad (3)$$

$$\text{where } \mathcal{E}_{\mathbf{M}_s, \mathbf{M}_t}(\mathbf{T}) \overset{\text{def.}}{=} \sum_{i,i',j,j'} \ell(M^s_{i,j}, M^t_{i',j'}) T_{i,i'} T_{j,j'},$$

where $M^s_{i,j}$ (*resp.*, $M^t_{i',j'}$) is the (i,j)-th (*resp.*, (i',j')-th) element of $\mathbf{M}_s$ (*resp.*, $\mathbf{M}_t$). The loss function $\ell(M^s_{i,j}, M^t_{i',j'})$ measures the difference between $M^s_{i,j}$ and $M^t_{i',j'}$ and is defined by $\ell(M^s_{i,j}, M^t_{i',j'}) = (M^s_{i,j} - M^t_{i',j'})^2$. The entropic regularization term is used to induce a smoother solution. Here we adopt the linear kernel matrix to construct the metric matrices, *i.e.*, $\mathbf{M}_s = \mathbf{X}_s \mathbf{X}_s^\top$ and $\mathbf{M}_t = \mathbf{X}_t \mathbf{X}_t^\top$.

Let the transported source samples in the target domain be $\tilde{\mathbf{X}}_s$, and the transported metric matrix be $\tilde{\mathbf{M}}_s = \tilde{\mathbf{X}}_s \tilde{\mathbf{X}}_s^\top$. After obtaining the transport matrix $\mathbf{T}$, $\tilde{\mathbf{X}}_s$ can be obtained by the GW barycenter [Peyré *et al.*, 2016], which is given as follows

$$\tilde{\mathbf{X}}_s = n_s \mathbf{T} \mathbf{X}_t. \quad (4)$$

Note that the size of the transport matrix $\mathbf{T}$ is $n_s \times n_t$. As a result, $\tilde{\mathbf{X}}_s$ have the same numbers of features to that of target samples. Although the transportation is linear, the feature transformation induced by $\mathbf{T}$ is usually highly nonlinear [Courty *et al.*, 2017]. As a result, the EGW potentially possesses a higher capacity compared with conventional HDA methods such as [Li *et al.*, 2014; Yan *et al.*, 2017b]. After learning the transportation, a classifier can be trained using the labeled target samples and transported labeled source samples, and then be applied to the unlabeled target samples for classification.

3.3 Semi-supervised Entropic Gromov-Wasserstein Discrepancy

A potential drawback when exploiting the entropic GW discrepancy for the HDA problem is that it does not exploit the label information in source and target domains. While it has shown superior performance in 3D object matching, the training samples in real-world tasks are more noisy and diverse. Thus, it is more desirable to further incorporate the label information to guide the learning of the transport matrix $\mathbf{T}$.

To this end, we propose a semi-supervised entropic Gromov-Wasserstein discrepancy (SGW) to seek for the optimal transport satisfying label consistency constraints in the target domain. Motivated by this, we propose the following learning problem to achieve the objective of label consistency based on both labeled and unlabeled training data,

$$\min_{\mathbf{T} \in \mathcal{T}} \mathcal{L}(\mathbf{T}) \overset{\text{def.}}{=} \mathcal{E}(\mathbf{T}) - \epsilon H(\mathbf{T}) + \lambda \Omega_l(\mathbf{T}) + \gamma \Omega_u(\mathbf{T}), \quad (5)$$

where $\mathcal{E}(\mathbf{T})$ is the abbreviation of $\mathcal{E}_{\mathbf{M}_s, \mathbf{M}_t}(\mathbf{T})$, $\Omega_l(\mathbf{T})$ is designed for labeled target samples, and $\Omega_u(\mathbf{T})$ is for unlabeled target samples. Next, we present the technical details regarding these two terms.

3.4 Conditional Distribution Matching

The solution to the EGW problem is able to adapt the marginal distributions of the target and transported source samples. To further match the conditional distributions of them, we enforce the centroids of target and transported source samples with the same label to approach each other. To this end, we minimize $\Omega_l(\mathbf{T})$ to make the data with the same label distribute close. Let $\tilde{\mathbf{X}}_s = n_s \mathbf{T} \mathbf{X}_t = [\tilde{\mathbf{x}}^s_1, \ldots, \tilde{\mathbf{x}}^s_{n_s}]^\top$ be the transported source instrances, $\Omega_l(\mathbf{T})$ is defined as follows:

$$\begin{aligned} \Omega_l(\mathbf{T}) &= \sum_{k=1}^{K} \Big|\Big| \frac{1}{n^s_k} \sum_{i=1}^{n^s_k} \tilde{\mathbf{x}}^s_{ik} - \frac{1}{n^l_k} \sum_{i=1}^{n^l_k} \mathbf{x}^l_{ik} \Big|\Big|_2^2 \\ &= \Big|\Big| n_s \mathbf{P} \mathbf{T} \mathbf{X}_t - \mathbf{Q} \mathbf{X}_l \Big|\Big|_F^2, \end{aligned} \quad (6)$$

where n^s_k and n^l_k are the number of labeled source and target samples with label k, $\tilde{\mathbf{x}}^s_{ik}$ and $\mathbf{x}^l_{ik}$ are samples with label k. $\mathbf{P} \in \mathbb{R}^{K \times n_s}$ and $\mathbf{Q} \in \mathbb{R}^{K \times n_l}$ are label indicator matrices and are constructed by the following equations,

$$P_{k,i} = \begin{cases} 1/n^s_k & \text{if } y^s_i = k, \\ 0 & \text{otherwise;} \end{cases} \quad (7)$$

$$Q_{k,i} = \begin{cases} 1/n^l_k & \text{if } y^l_i = k, \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

$\Omega_l(\mathbf{T})$ is identical to the conditional maximum mean discrepancy when using the linear kernel, which measures the conditional distribution discrepancy between two sets of samples [Long *et al.*, 2014; Tsai *et al.*, 2016].

3.5 Group Entropic Regularization

Recall that the transport matrix $\mathbf{T}$ models how likely each source sample will be transported to each target sample,

which is also referred to as *mass*. Intuitively, given a target sample, most of the mass of it should be transported from the source samples with the same label. Therefore, we divide the source samples into different groups according to their labels, and for the j-th column of the joint distribution $\mathbf{T}_{\cdot,j}$, we design a group entropic regularization term to make the probability distribute on one group. Specifically, let $\mathcal{I}_k$ be the indices of the source samples with label k, the probability $T_{\mathcal{I}_k,j}$ is the sum of the probabilities from the source samples with label k to j-th target sample, *i.e.*, $T_{\mathcal{I}_k,j} = \sum_{i\in\mathcal{I}_k} T_{i,j}$, the regularization term is defined as

$$\Omega_u(\mathbf{T}) = \sum_{j=1}^{n_t} H(T_{\mathcal{I}_1,j}, \ldots, T_{\mathcal{I}_K,j}). \tag{9}$$

By minimizing the above group entropic regularization term, the probability for one target sample is concentrated on one group of source samples with the same label.

4 Optimization

Problem (5) is non-convex with the equality constraints, thus is difficult to solve. To efficiently solve it, we apply a projected gradient descent algorithm according to the Kullback-Leibler divergence [Benamou *et al.*, 2015; Peyré *et al.*, 2016]. Specifically, at the τ-th iteration, $\mathbf{T}_\tau$ is firstly updated by the exponentiated gradient method as follows:

$$\hat{\mathbf{T}}_\tau := \mathbf{T}_\tau \odot \exp\big(-\alpha\nabla\mathcal{L}(\mathbf{T}_\tau)\big), \tag{10}$$

where $\alpha > 0$ is a step size. And then $\hat{\mathbf{T}}_\tau$ is projected into the definition domain $\mathcal{T}$ with the Kullback-Leibler metric

$$\mathbf{T}_{\tau+1} := \Pi_{\mathcal{T}}^{\mathrm{KL}}(\hat{\mathbf{T}}_\tau) = \arg\min_{\mathbf{T}'\in\mathcal{T}} \mathrm{KL}(\mathbf{T}'|\hat{\mathbf{T}}_\tau). \tag{11}$$

From [Benamou *et al.*, 2015], the projection operation in Eq. (11) can be rewritten as the following regularized optimal transport problem

$$\mathbf{T}_{\tau+1} := \Pi_{\mathcal{T}}^{\mathrm{KL}}(\hat{\mathbf{T}}_\tau) = \arg\min_{\mathbf{T}'\in\mathcal{T}} \langle -\epsilon\log(\hat{\mathbf{T}}_\tau), \mathbf{T}'\rangle - \epsilon H(\mathbf{T}'), \tag{12}$$

which can be solved by the Sinkhorn's fixed point algorithm [Cuturi, 2013]. Specifically, the transport cost matrix $-\epsilon\log(\hat{\mathbf{T}}_\tau)$ can be simplified as

$$\begin{aligned} -\epsilon\log(\hat{\mathbf{T}}_\tau) &= -\epsilon\log\Big(\mathbf{T}_\tau \odot \exp\big(-\alpha\nabla\mathcal{L}(\mathbf{T}_\tau)\big)\Big) \\ &= \nabla\mathcal{E}(\mathbf{T}_\tau) + \nabla\Omega(\mathbf{T}_\tau), \end{aligned} \tag{13}$$

where we set $\epsilon\alpha = 1$, and $\Omega(\mathbf{T}_\tau) = \lambda\Omega_l(\mathbf{T}_\tau) + \gamma\Omega_u(\mathbf{T}_\tau)$. The solution to Problem (12) is

$$\mathbf{T}_{\tau+1} = \mathrm{diag}(\mathbf{u})\ \boldsymbol{\Theta}\ \mathrm{diag}(\mathbf{v}), \tag{14}$$

where the matrix $\boldsymbol{\Theta}$ is constructed by

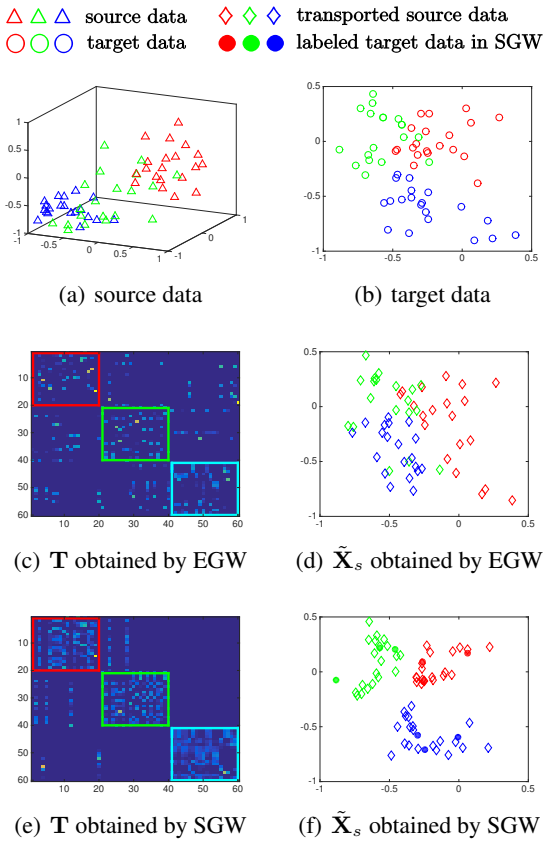
$$\boldsymbol{\Theta} = \exp\big(-\frac{1}{\epsilon}\big(\nabla\mathcal{E}(\mathbf{T}_\tau) + \nabla\Omega(\mathbf{T}_\tau)\big)\big), \tag{15}$$

and the vectors $(\mathbf{u}, \mathbf{v})$ are computed using Sinkhorn's fixed point iterations as follows.

$$\mathbf{u} := \frac{\mathbf{p}_s}{\boldsymbol{\Theta}\mathbf{v}}, \quad \mathbf{v} := \frac{\mathbf{p}_t}{\boldsymbol{\Theta}^\top\mathbf{u}}, \tag{16}$$

where the division operations are performed element-wisely.

Algorithm 1 summarizes the main steps of the proposed SGW algorithm.



(a) source data (b) target data

(c) $\mathbf{T}$ obtained by EGW (d) $\tilde{\mathbf{X}}_s$ obtained by EGW

(e) $\mathbf{T}$ obtained by SGW (f) $\tilde{\mathbf{X}}_s$ obtained by SGW

Figure 1: Simulated results on Wine dataset. (a) and (b) are the source and target data points, respectively. (c) and (d) are the transport matrix and transported source samples obtained by EGW. (e) and (f) are the transported matrix and transported source samples obtained by SGW.

5 Experiment

5.1 Synthetic Data

To better illustrate our proposed approach, we use synthetic data to visualize the transport matrices and the transported source data obtained by EGW and SGW. The synthetic data are generated from Wine dataset [1], which includes 178 samples with three classes. To generate heterogeneous source and target samples, we randomly pick up three and two features as the source and target features, respectively.

Figures 1(a) and 1(b) show the source and target data, where three colors correspond to three classes. The transport matrices obtained by EGW and SGW are visualized in Figures 1(c) and 1(e), respectively, in which a lighter point is a larger value indicating the corresponding joint distribution of the source and target samples is higher. For better observability, we order samples according to their labels to make the samples with the same label gather together. Therefore, in a label consistency preserving transport matrix, the larger values must be in the colored square frames. The transported source data in the target feature space are shown in Figures 1(d) and 1(f), respectively.

[1] https://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/

S → T	SVM	CCA+OT	EGW	SHFA	CDLS	DCA	SGW
DeCAF$_6$ → SURF							
A → C		30.58 ± 3.31	28.13 ± 3.76	30.42 ± 3.04	31.94 ± 3.38	32.35 ± 3.48	$\mathbf{34.54 \pm 1.42}$
D → C	28.42 ± 2.68	31.37 ± 2.81	30.64 ± 1.45	31.58 ± 2.30	32.73 ± 2.24	32.75 ± 2.27	$\mathbf{34.72 \pm 0.93}$
W → C		31.68 ± 2.47	27.88 ± 2.96	30.81 ± 2.47	33.12 ± 2.26	33.25 ± 3.05	$\mathbf{35.03 \pm 1.61}$
Average	28.42	31.21	28.88	30.94	32.60	32.78	**34.76**
SURF → DeCAF$_6$							
C → A	79.63 ± 3.42	84.06 ± 2.01	85.35 ± 2.85	85.33 ± 1.65	83.70 ± 2.37	89.16 ± 1.62	$\mathbf{92.15 \pm 0.39}$
C → D	87.29 ± 5.98	90.75 ± 3.45	89.72 ± 4.36	92.52 ± 4.29	91.31 ± 3.15	92.90 ± 4.41	$\mathbf{95.14 \pm 0.59}$
C → W	86.86 ± 2.76	88.20 ± 3.60	88.04 ± 3.92	88.57 ± 1.22	88.29 ± 4.10	89.22 ± 3.03	$\mathbf{95.39 \pm 1.11}$
Average	84.59	87.67	87.70	88.81	87.76	90.43	**94.23**

Table 1: Results on the Office-Caltech dataset using 3 labeled target samples

Algorithm 1 Semi-supervised Entropic Gromov-Wasserstein.

Initialize: $\mathbf{T}_1 = \mathbf{p}_s \mathbf{p}_t^\top$, $\tau = 1$.
1: **repeat**
2: Compute $\tilde{\mathbf{X}}_s$ by Eq. (4).
3: Construct $\mathbf{\Theta}$ based on $\mathbf{T}_\tau$ and Eq. (15).
4: Obtain $(\mathbf{u}, \mathbf{v})$ by the fixed point iterations in Eq. (16).
5: Update $\mathbf{T}_{\tau+1}$ by Eq. (14)
6: $\tau := \tau + 1$.
7: **until** Convergence.
8: Train a classifier on $(\tilde{\mathbf{X}}_s, \mathbf{Y}_s)$ and $(\mathbf{X}_l, \mathbf{Y}_l)$.
9: Predict for the unlabeled target data $\mathbf{X}_u$.

From the results of EGW in Figures 1(c) and 1(d), we observe that although the transported source samples follow a similar distribution to the target samples, some source samples are transported to the mismatching regions with different labels in the target domain. For SGW, three labeled target samples are used for training, and the results are shown in Figures 1(e) and 1(f). We observe that most source samples are transported to the regions with the same labels in the target domain, which verifies the effect of $\Omega_l(\mathbf{T})$ in Eq. (6). In addition, in Figure 1(e), one target sample is usually transported from a group of source samples with the same label, which demonstrates the effect of our proposed group entropic regularization in Eq. (9).

5.2 Real-World Datasets

To further validate the effectiveness of our proposed approach, we conduct experiments on two benchmark real-world datasets for object recognition and text classification, respectively.

Object Recognition: We use Office and Caltech-256 datasets for the object recognition task. Office [Saenko *et al.*, 2010] includes images with 31 classes from three domains: amazon (A), dslr (D) and webcam (W). We use the publicly available SURF [Bay *et al.*, 2006] and DeCAF$_6$ [Donahue *et al.*, 2014] features, and the dimensions of SURF and DeCAF$_6$ features are 800 and 4096, respectively. Caltech-256 (C) [Griffin *et al.*, 2007] includes images with 256 classes, and the same 800-d SURF features are publicly available. Following [Tsai *et al.*, 2016; Yan *et al.*, 2017b], we use the 10 overlapping classes between two datasets to construct classification tasks.

Text Classification: The text classification task is conducted on the Reuters multilingual dataset [Amini *et al.*, 2009], which includes six classes and five languages, *i.e.*, English, French, German, Italian and Spanish. Each language is taken as a domain. Similar to the experimental setup in [Li *et al.*, 2014; Tsai *et al.*, 2016], The documents are represented by TF-IDF features and are preprocessed by PCA with 60% energy preserved.

5.3 Baseline Methods

We compare with SVM [Chang and Lin, 2011], CCA+OT, EGW, SHFA [Li *et al.*, 2014], CDLS [Tsai *et al.*, 2016] and DCA [Yan *et al.*, 2017b]. Among these methods, SVM is a simple baseline without considering source domain data. CCA+OT firstly performs CCA on source and target data, and then transports source data into the target domain based on a classical optimal transport method. EGW learns the optimal transport matrix by solving the EGW problem, and an SVM classifier is trained on the transported source samples. SHFA, CDLS and DCA are three state-of-the-art HDA methods, which are performed in a semi-supervised HDA fashion, where some unlabeled target samples are used for training.

For simplicity and fair comparison, we set the trade-off parameter of SVM to $C = 1$ for all the methods and tasks. The parameters of SGW are empirically set to $\epsilon = 0.01$, $\lambda = 1$ and $\gamma = 1$, and the sensitivity study is provided in Section 5.6. For the baseline methods, we follow the previous work [Li *et al.*, 2014] to search the parameters in the spaces recommended by the original papers and report their best results. We repeatedly conduct experiments 10 trials and report the average classification accuracies.

5.4 Results on Object Recognition Tasks

Office includes three domains (*i.e.*, A, D and W), which are represented as two types of features (*i.e.*, SURF and DeCAF$_6$), and Caltech-256 dataset includes one domain (*i.e.*, C) with SURF features. We take C as the source domain, and a domain in Office as the target domain, and vice versa.

We randomly choose 3 target samples per class as the labeled target data, and the rest target samples as the test data. 20 labeled source samples are used for training for the source domains A, W and C, and 5 labeled source samples are used for training for the source domain D, since the number of samples in D is much smaller.

Table 1 presents the results of the Office-Caltech dataset. SGW achieves the best results, which demonstrates the ef-

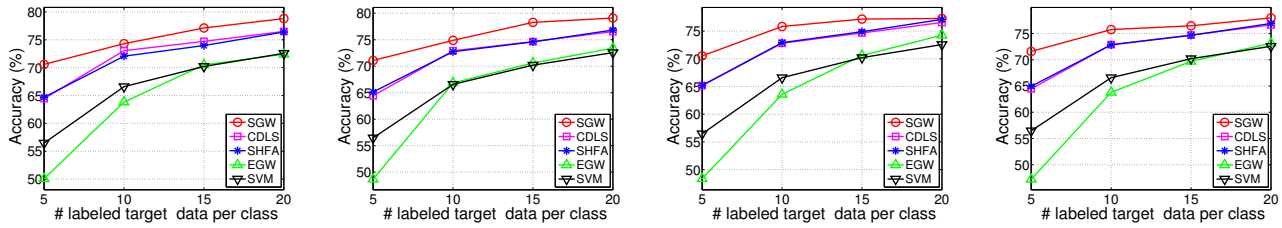


Figure 2: Classification results on the text classification tasks. Spanish is taken as the target domain, and the source domains are selected from (a) English, (b) French, (c) German, and (d) Italian, respectively.

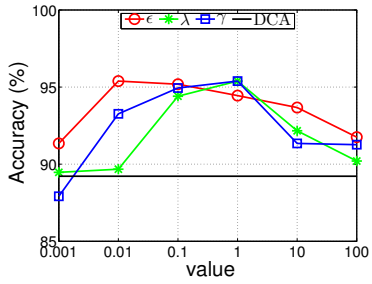


Figure 3: Sensitivity study of SGW on the task C(SURF) $\rightarrow$ W(DeCAF$_6$) *w.r.t.* parameters ϵ, λ and γ.

fectiveness of our proposed method. The domain adaptation methods CCA+OT, SHFA, CDLS and DCA outperform SVM, which verifies that leveraging knowledge from heterogeneous source domain is beneficial for the target task. Compared with SVM, EGW achieves improvements in most cases, which shows that the EGW discrepancy is helpful for matching the data structure between two heterogeneous domains. However, there are still a few cases that EGW is even worse than SVM. A possible reason is that the samples from different categories could be mixed up after transportation since the label information is not employed in EGW. Our proposed SGW obtains consistent better performance than the baseline methods. We attribute this to the effective use of label information for guiding the optimal transport.

5.5 Results on Text Classification Tasks

Similar to the settings in [Li *et al.*, 2014; Tsai *et al.*, 2016], we take Spanish as the target domain, and the other four languages as the source domain, respectively. We randomly choose 100 source samples per class as the labeled source data. For the target domain, $\{5, 10, 15, 20\}$ samples are used as labeled training samples, and the remaining samples are used as test data, among which 3,000 samples are unlabeled training data for semi-supervised learning.

Figure 2 shows the results on the text classification tasks *w.r.t.* different numbers of labeled target samples. We observe that the performance of all the methods increases when using more labeled target samples for training, which validates the effect of adding more labeled target data. Our SGW approach consistently outperforms all the baselines, which demonstrates the effectiveness of our proposed method for heterogeneous domain adaptation.

	SVM	SHFA	CDLS	DCA	CCA+OT	EGW	SGW
Time (s)	0.22	21.53	34.14	83.63	5.47	2.06	2.17

Table 2: Running time on the task C(SURF) $\rightarrow$ W(DeCAF$_6$).

5.6 Sensitivity Study

We take the task C(SURF) $\rightarrow$ W(DeCAF$_6$) as an example to study the parameter sensitivity of SGW. We vary the trade-off parameters ϵ, λ and γ in the search space $\{0.001, 0.01, 0.1, 1, 10, 100\}$, and plot the results in Figure 3. DCA achieves the second best result on this task, thus is taken as the reference. In most cases, SGW outperforms DCA, which demonstrates that the performance of SGW is stable to the trade-off parameters in the certain range. We have similar observations on other tasks.

5.7 Running Time Results

We take C(SURF) $\rightarrow$ W(DeCAF$_6$) as a representative task to evaluate the efficiencies of all the methods. The experiments are performed on a workstation with Xeon 3.40 GHz CPU and 16 GB of RAM. Table 2 presents the running time results. SVM only involves target data, thus achieves the shortest running time. SHFA, CDLS and DCA learn feature transformations and classifiers, and have to solve quadratic programming problems with heavy matrix computations. As a result, the running times of them are much longer than SVM. EGW and SGW usually converge to a good solution within a few iterations. Therefore, the running times of EGW and SGW are less than those of SHFA, CDLS and DCA.

6 Conclusion

In this paper, we propose a new algorithm using optimal transport for heterogeneous domain adaptation. We learn an optimal transport matrix to transport labeled source samples into the target domain. To incorporate label information to guide the learning of optimal transport, we propose a semi-supervised entropic Gromov-Wasserstein discrepancy, which remains the metric matrices on source data before and after transportation, and makes the labeled target and transported source samples follow similar conditional distributions. We conduct extensive experiments on both synthetic and real-world datasets, and the results demonstrate the effectiveness and efficiency of our proposed method.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) 61502177 and 61602185, and Recruitment Program for Young Professionals, and Guangdong Provincial Scientific and Technological funds 2017B090901008, 2017A010101011, 2017B090910005, and Fundamental Research Funds for the Central Universities D2172500, D2172480, and Pearl River S&T Nova Program of Guangzhou 201806010081 and CCF-Tencent Open Research Fund RAGR20170105.

References

[Amini *et al.*, 2009] Massih Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. In *NIPS*, pages 28–36, 2009.

[Bay *et al.*, 2006] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, pages 404–417, 2006.

[Benamou *et al.*, 2015] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *TIST*, 2(3):27, 2011.

[Courty *et al.*, 2017] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *TPAMI*, 39(9):1853–1865, 2017.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, pages 2292–2300, 2013.

[Donahue *et al.*, 2014] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, volume 32, pages 647–655, 2014.

[Duan *et al.*, 2012] Lixin Duan, Dong Xu, and Ivor W Tsang. Learning with augmented features for heterogeneous domain adaptation. In *ICML*, pages 711–718, 2012.

[Griffin *et al.*, 2007] Gregory Griffin, Alex Holub, and Pietro Perona. *Caltech-256 object category dataset*. California Institute of Technology, 2007.

[Li *et al.*, 2014] Wen Li, Lixin Duan, Dong Xu, and Ivor W Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *TPAMI*, 36(6):1134–1148, 2014.

[Liu *et al.*, 2017] Tongliang Liu, Qiang Yang, and Dacheng Tao. Understanding how feature structure transfers in transfer learning. In *IJCAI*, pages 2365–2371, 2017.

[Long *et al.*, 2014] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer joint matching for unsupervised domain adaptation. In *CVPR*, pages 1410–1417, 2014.

[Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015.

[Luo *et al.*, 2017] Yong Luo, Yonggang Wen, Tongliang Liu, and Dacheng Tao. General heterogeneous transfer distance metric learning via knowledge fragments transfer. In *IJCAI*, pages 2450–2456, 2017.

[Moon and Carbonell, 2017] Seungwhan Moon and Jaime Carbonell. Completely heterogeneous transfer learning with attention-what and what not to transfer. In *IJCAI*, pages 2508–2514, 2017.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *TKDE*, 22(10):1345–1359, 2010.

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

[Patel *et al.*, 2015] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015.

[Perrot *et al.*, 2016] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *NIPS*, pages 4197–4205, 2016.

[Peyré and Cuturi, 2017] Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2017.

[Peyré *et al.*, 2016] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *ICML*, pages 2664–2672, 2016.

[Redko *et al.*, 2017] Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *ECML-PKDD*, pages 737–753, 2017.

[Saenko *et al.*, 2010] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010.

[Shao *et al.*, 2015] Ling Shao, Fan Zhu, and Xuelong Li. Transfer learning for visual categorization: A survey. *TNNLS*, 26(5):1019–1034, 2015.

[Tan *et al.*, 2015] Ben Tan, Yangqiu Song, Erheng Zhong, and Qiang Yang. Transitive transfer learning. In *KDD*, pages 1155–1164, 2015.

[Tsai *et al.*, 2016] Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Learning cross-domain landmarks for heterogeneous domain adaptation. In *CVPR*, pages 5081–5090, 2016.

[Wu *et al.*, 2014] Qingyao Wu, Michael K. Ng, and Yunming Ye. Cotransfer learning using coupled markov chains with restart. *IEEE Intelligent Systems*, 29(4):26 – 33, 2014.

[Yan *et al.*, 2017a] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *CVPR*, pages 2272–2281, 2017.

[Yan *et al.*, 2017b] Yuguang Yan, Wen Li, Michael Ng, Mingkui Tan, Hanrui Wu, Huaqing Min, and Qingyao Wu. Learning discriminative correlation subspace for heterogeneous domain adaptation. In *IJCAI*, pages 3252–3258, 2017.

[Yan *et al.*, 2017c] Yuguang Yan, Qingyao Wu, Mingkui Tan, Michael K Ng, Huaqing Min, and Ivor W Tsang. Online heterogeneous transfer by hedge ensemble of offline and online decisions. *TNNLS*, pages 1–12, 2017.

[Zhou *et al.*, 2014] Joey Tianyi Zhou, Sinno Jialin Pan, Ivor W Tsang, and Yan Yan. Hybrid heterogeneous transfer learning through deep learning. In *AAAI*, pages 2213–2220, 2014.

[Zhou *et al.*, 2016] Joey Tianyi Zhou, Sinno Jialin Pan, Ivor W Tsang, and Shen-Shyang Ho. Transfer learning for cross-language text categorization through active correspondences construction. In *AAAI*, pages 2400–2406, 2016.