Machine Learning (2022) 111:4159–4182
https://doi.org/10.1007/s10994-022-06231-7

Hierarchical optimal transport for unsupervised domain adaptation

Mourad El Hamri[1,2] · **Younès Bennani**[1,2] · **Issam Falih**[2,3]

Received: 21 December 2021 / Revised: 27 July 2022 / Accepted: 9 August 2022 /
Published online: 30 September 2022

Abstract
In this paper, we propose a novel approach for unsupervised domain adaptation that relates notions of optimal transport, learning probability measures, and unsupervised learning. The proposed approach, HOT-DA, is based on a hierarchical formulation of optimal transport that leverages beyond the geometrical information captured by the ground metric, richer structural information in the source and target domains. The additional information in the labeled source domain is formed instinctively by grouping samples into structures according to their class labels. While exploring hidden structures in the unlabeled target domain is reduced to the problem of learning probability measures through Wasserstein barycenter, which we prove to be equivalent to spectral clustering. Experiments show the superiority of the proposed approach over state-of-the-art across a range of domain adaptation problems including inter-twinning moons dataset, Digits, Office-Caltech, and Office-Home. Experiments also show the robustness of our model against structure imbalance. We make our code publicly available.

Keywords Optimal transport · Domain adaptation · Learning probability measures · Unsupervised learning

Editors:Krzysztof Dembczynski and Emilie Devijver.

✉ Mourad El Hamri
mourad.elhamri@sorbonne-paris-nord.fr

Younès Bennani
younes.bennani@sorbonne-paris-nord.fr

Issam Falih
issam.falih@uca.fr

[1] LIPN CNRS UMR 7030, Université Sorbonne Paris Nord, Villetaneuse, France

[2] La Maison des Sciences Numériques, Saint-Denis, France

[3] LIMOS CNRS UMR 6158, Université Clermont Auvergne, Clermont-Ferrand, France

1 Introduction

Supervised learning is arguably the most widespread task of machine learning and has enjoyed much success on a broad spectrum of application domains (Kotsiantis et al., 2007). However, most supervised learning methods are built on the crucial assumption that training and test data are drawn from the same probability distribution (Pan and Yang, 2009). In real-world applications, this hypothesis is usually violated due to several application-dependent reasons: in computer vision, the presence or absence of backgrounds, the variation of acquisition devices, or the change of lighting conditions introduce non-negligible discrepancies in data distributions (Saenko et al., 2010), in product reviews classification, the drifts observed in the word distributions are caused by the difference of product category and the changes in word frequencies (Blitzer et al., 2007). These distributional shifts will be likely to degrade significantly the generalization ability of supervised learning models. While manual labeling may appear like a feasible solution, such an approach is unreasonable in practice, since it is often prohibitively expensive to collect from scratch a new large high quality labeled dataset with the same distribution as the test data, due to lack of time, resources, or other factors, and it would be an immense waste to totally reject the available knowledge on a different, yet related labeled training set. Such a challenging situation has promoted the emergence of domain adaptation (Redko et al., 2019), a subfield of statistical learning theory (Vapnik, 2013), that takes into account the distributional shift between training and test data, and in which the training set and test set distributions are respectively called source and target domains. There are two variants of domain adaptation, depending on the availability of a small amount of labeled data in the target domain (semi-supervised domain adaptation) or not (unsupervised domain adaptation). This paper deals with the challenging setting of unsupervised domain adaptation.

Since the launching of domain adaptation theory, a large panoply of algorithms was proposed to deal with its unsupervised variant, and they can be roughly divided into shallow (Kouw and Loog, 2019) and deep (Wilson and Cook, 2020) approaches. Most of shallow algorithms try to solve the unsupervised domain adaptation problem in two steps by first aligning the source and target domains to make them indiscernible, which then allows to apply traditional supervised methods on the transformed data. Such an alignment is typically accomplished through sample-based approaches which focus on correcting biases in the sampling procedure (Shimodaira, 2000; Sugiyama et al., 2007) or feature-based approaches which focus on learning domain-invariant representations (Pan et al., 2010) and finding subspace mappings (Gong et al., 2012; Fernando et al., 2013). Deep domain adaptation algorithms have also gained a renewed interest due to their feature extraction ability to learn more abstract and robust representations that are both semantically meaningful and domain invariant. Ganin et al. (2016) is one of the most popular deep adaptative networks, which is based on the adversarial training procedure (Goodfellow et al., 2014) and directly derived from the seminal theoretical contribution in Ben-David et al. (2006), its main idea is to embed domain adaptation into the representation learning process, so that the final classification decisions are made based on features that are both discriminative and invariant to domain changes. Later on, the work of Zhang et al. (2019) provided margin-aware generalization bounds, which can also be transformed into an adversarial learning algorithm for domain adaptation.

More recent advances in domain adaptation are due to the theory of optimal transport (Villani, 2009), which allows to learn explicitly the least cost transformation of the source distribution into the target one. This idea was first investigated in the work of

Courty et al. (2016), where authors have successfully cast the domain adaptation problem into an optimal transport one to match the shifted marginal distributions of the two domains, which then allows to learn a classifier on the transported data. Since then, several optimal transport based domain adaptation methods have emerged. In Courty et al. (2017), authors proposed to avoid the two-steps adaptation procedure, by aligning the joint distributions using a coupling accounting for the marginals and the class-conditional distributions shift jointly. Authors in Redko et al. (2019) performed multi-source domain adaptation under the target shift assumption, by learning simultaneously the class probabilities of the unlabeled target samples and the optimal transport plan allowing to align several probability distributions. The recent work of Dhouib et al. (2020) derived an efficient optimal transport based adversarial approach from a bound on the target margin violation rate. Finally, several deep domain adaptation algorithms based on optimal transport were proposed in Damodaran et al. (2018), Shen et al. (2018), Chen et al. (2018), Xu et al. (2020), Li et al. (2020) to name a few.

A common denominator of these approaches is their ability to capture the underlying geometry of the data by relying on the cost function that reflects the metric of the input space. However, these optimal transport based methods can benefit from not relying solely on such rudimentary geometrical information, since there is further important structural information that remains uncaptured directly from the ground metric, e.g., the local consistency induced by class labels in the source. The exploitation of this structural information can elicit some desired properties in domain adaptation like preserving compact classes during the transportation. It is, moreover, what led authors in Courty et al. (2016) to propose the inclusion of this structural information by adding a group-norm regularizer. Such structures, however, could not be induced directly by the standard formulation of optimal transport. To the best of our knowledge, Alvarez-Melis et al. (2018) is the only work that has attempted to incorporate structural information directly into the optimal transport problem without the need to add a regularization term. This approach developed a nonlinear generalization of discrete optimal transport based on submodular functions. However, the application of this method in domain adaptation only takes into account the available structures in the labeled source domain, by partitioning samples according to their class labels, while every target sample forms its own cluster. Nonetheless, richer structures in the target domain can be easily captured differently, e.g., by grouping, and the incorporation of such target structures directly into the optimal transport formulation can lead in our view to a significant improvement in the performance of domain adaptation algorithms.

Contributions and outline of the paper: In this paper, we address the existing limitations of the target-structure-agnostic algorithms mentioned above by proposing a principally new approach based on hierarchical optimal transport (Schmitzer and Schnörr, 2013). Hierarchical optimal transport is an effective and efficient paradigm to induce structural information into the transportation procedure. It has been recently used for different tasks such as multi-level clustering (Ho et al., 2017), multimodal distribution alignment (Lee et al., 2019), document representation (Yurochkin et al., 2019) and semi-supervised learning (Taherkhani et al., 2020). The relevance of this paradigm for domain adaptation is illustrated in Fig. 1, where we show that the structure-agnostic Reg-OT (Cuturi, 2013) and target-structure-agnostic OT-GL (Courty et al., 2016) algorithms fail to always restrict the transportation of mass across instances of different structures, whereas, our Hierarchical Optimal Transport for Domain Adaptation (HOT-DA) model manages to do it correctly by leveraging the source and target structures simultaneously, which will subsequently lead to a better adaptation.

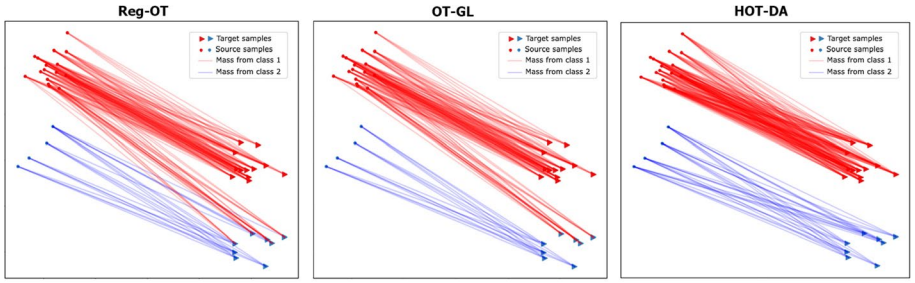


Fig. 1 Illustration of the transportation obtained with structure-agnostic Reg-OT (Cuturi, 2013) and target-structure-agnostic OT-GL (Courty et al., 2016) methods, and our proposed algorithm HOT-DA

To the best of our knowledge, the proposed approach is the first hierarchical optimal transport method for unsupervised domain adaptation, and the first work to shed light on the connection between spectral clustering and Wasserstein barycenter.

The rest of this paper is organized as follows: in Sect. 2, we present a brief overview of unsupervised domain adaptation setup. In Sect. 3, we detail the optimal transport problem and its hierarchical formulation, then in Sect. 4, we elaborate the proposed approach HOT-DA. In Sect. 5, we evaluate our algorithm on a toy dataset and three benchmark visual adaptation problems, and we study the relevance of Wasserstein-Spectral clustering to HOT-DA as well as the sensitivity of our approach to unbalanced structures. Finally, we conclude in Sect. 6.

2 Unsupervised domain adaptation

Let $\mathcal{X} = \mathbb{R}^d$ be an input space, $\mathcal{Y} = \{c_1, \dots, c_k\}$ a discrete label space consisting of k classes, $\mathcal{S}$ and $\mathcal{T}$ two different probability distributions over $\mathcal{X} \times \mathcal{Y}$ called respectively the source and target domains. We have access to a set $S = \{(x_i, y_i)\}_{i=1}^n$ of n labeled source samples drawn i.i.d. from the joint distribution $\mathcal{S}$ and a set $T = \{x_j\}_{j=1}^m$ of m unlabeled target samples drawn i.i.d. from the marginal distribution $\mathcal{T}_\mathcal{X}$, of the joint distribution $\mathcal{T}$ over $\mathcal{X}$, more formally:

$$S = \{(x_i, y_i)\}_{i=1}^n \sim (\mathcal{S})^n, \quad T = \{x_j\}_{j=1}^m \sim (\mathcal{T}_\mathcal{X})^m$$

The aim of unsupervised domain adaptation algorithms is to infer a classifier $\eta : \mathcal{X} \to \mathcal{Y}$ with a low target risk:

$$\mathcal{R}_\mathcal{T}(\eta) = \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{T}}(\eta(x) \neq y)$$

under the distributional shift assumption $\mathcal{S} \neq \mathcal{T}$, while having no information about the labels $\{y_j\}_{j=1}^m$ of the target set T, other than the fact that T and S share the same label space $\mathcal{Y}$. In the rest, we design by the source domain interchangeably the distribution $\mathcal{S}$ and the labeled set S, and by the target domain, the distribution $\mathcal{T}$ and the unlabeled set T.

3 Optimal transport

In this section, we present the key concepts of the optimal transport problem and its hierarchical formulation (Villani, 2009).

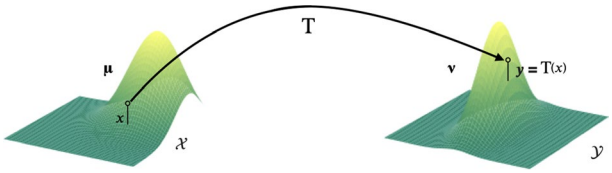


Fig. 2 Monge's problem: T is a transport map from $\mathcal{X}$ to $\mathcal{Y}$.

Optimal transport is a long-standing mathematical problem whose theory has matured over time. Its roots can be traced back to the 18$^{\text{th}}$ century, when the French mathematician Gaspard Monge introduced the following problem (Monge, 1781): Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be two probability spaces, $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ a positive cost function over $\mathcal{X} \times \mathcal{Y}$, which represents the work needed to move a unit of mass from $x \in \mathcal{X}$ to $y \in \mathcal{Y}$. The problem asks to find a measurable transport map $\mathrm{T} : \mathcal{X} \to \mathcal{Y}$ that transports the mass represented by the probability measure μ to the mass represented by the probability measure ν, while minimizing the total cost of this transportation:

$$(\mathcal{M}) \quad \inf_{\mathrm{T}} \{ \int_{\mathcal{X}} c(x, \mathrm{T}(x)) d\mu(x) | \mathrm{T}\#\mu = \nu \} \tag{1}$$

where $\mathrm{T}\#\mu$ stands for the image measure of μ by T. The problem of Monge $(\mathcal{M})$ depicted in Fig. 2 is quite difficult, since it is not symmetric, and may not admit a solution, it is the case when μ is a Dirac measure and ν is not.

A long period of sleep followed Monge's formulation until the convex relaxation of the Soviet mathematician Leonid Kantorovitch in the thick of World War II (Kantorovich, 1942). This relaxed formulation, known as the problem of Monge-Kantorovich $(\mathcal{MK})$ allows mass splitting and, in contrast to the formulation of Monge, it guarantees the existence of a solution under very general assumptions:

$$(\mathcal{MK}) \quad \inf_{\gamma} \{ \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\gamma(x, y) | \gamma \in \Pi(\mu, \nu) \} \tag{2}$$

where $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) | proj_{\mathcal{X}}\#\gamma = \mu, \, proj_{\mathcal{Y}}\#\gamma = \nu\}$ is the transport plans set, constituted of all joint probability measures γ on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν.

When $\mathcal{X} = \mathcal{Y}$ is a polish metric space endowed with a distance d, a natural choice is to use it as a cost function, e.g., $c(x, y) = d(x, y)^p$ for $p \in [1, +\infty[$. Then, the problem $(\mathcal{MK})$ induces a metric between probability measures over $\mathcal{X}$, called the p-Wasserstein distance (Santambrogio, 2015). The p-Wasserstein distance is defined in the following way, $\forall \mu, \nu \in \mathcal{P}(\mathcal{X})$:

$$\mathrm{W}_p(\mu, \nu) = (\inf_{\gamma \in \Pi(\mu, \nu)} \int_{\mathcal{X}^2} d^p(x, y) \, d\gamma(x, y))^{1/p} \tag{3}$$

In the discrete version of optimal transport, i.e., when the measures μ and ν are only available through discrete samples $X = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ and $Y = \{y_1, \ldots, y_m\} \subset \mathcal{Y}$, their empirical distributions can be expressed as $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$, where $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_m)$ are vectors in the probability simplex $\sum_n$ and $\sum_m$ respectively. The cost function only needs to be specified for every pair $(x_i, y_j)_{\substack{1 \le i \le n \\ 1 \le j \le m}} \in X \times Y$ yielding a cost matrix $C \in \mathcal{M}_{n \times m}(\mathbb{R}^+)$. The problem $(\mathcal{MK})$ becomes then a linear program (Bertsimas and Tsitsiklis, 1997) parametrized by the transportation polytope $U(a, b) = \{\gamma \in \mathcal{M}_{n \times m}(\mathbb{R}^+) | \gamma \mathbb{1}_m = a \text{ and } \gamma^{\mathbf{T}} \mathbb{1}_n = b\}$, which acts as a feasible set, and the

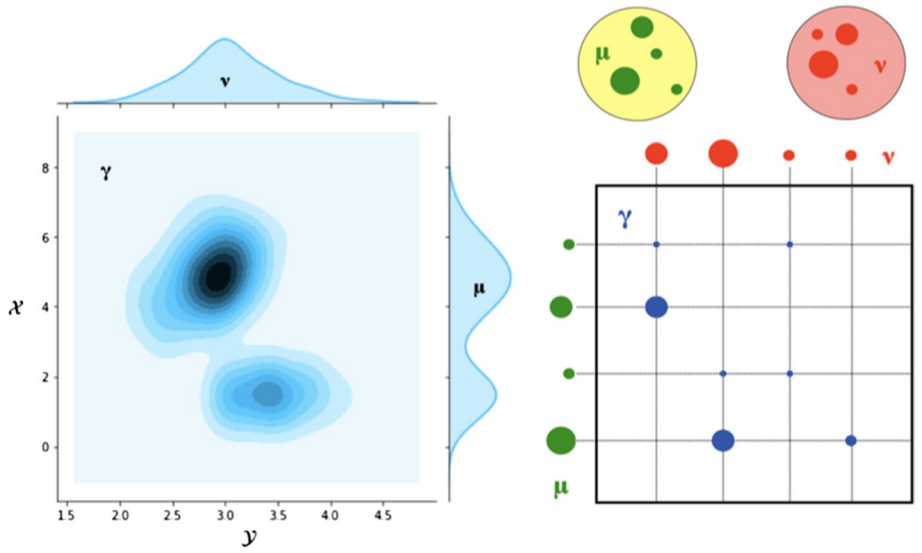


Fig. 3 Continuous Kantorovich's relaxation: the joint probability distribution γ is a transport plan between μ and ν (left). Discrete Kantorovich's relaxation: the positive entries of the discrete transport plan are displayed as blue disks with a radius proportional to the entry values (right) (Color figure online)

matrix C which acts as a cost parameter. Thus, solving this linear program consists in finding a plan γ^* that realizes:

$$(\mathcal{D}_{\mathcal{MK}}) \quad \min_{\gamma \in U(a,b)} \langle \gamma, C \rangle_F \tag{4}$$

where $\langle ., . \rangle_F$ is the Frobenius inner product. In this case, the p-Wasserstein distance can be inferred as follows: $\mathrm{W}_p^p(\mu, \nu) = \langle \gamma^*, C \rangle_F$. An illustration of the problem of Monge-Kantorovich in its continuous and discrete formulation is provided in Fig. 3.

A Wasserstein barycenter (Agueh and Carlier, 2011) of N measures $\{\nu_1, \ldots, \nu_N\}$ in $\mathcal{P}(\mathcal{X})$ can be defined as a minimizer of the following functional f over $\mathcal{P}(\mathcal{X})$:

$$f(\kappa) = \frac{1}{N} \sum_{i=1}^{N} \lambda_i \mathrm{W}_p^p(\kappa, \nu_i) \tag{5}$$

where λ_i are positive real numbers such that $\sum_{i=1}^{N} \lambda_i = 1$.

As stated above, discrete optimal transport is a linear program, and thus can be solved exactly in $\mathcal{O}(r^3 \log(r))$ where $r = \max(n, m)$, with the simplex algorithm or interior point methods (Pele and Werman, 2009), which is a heavy computational price tag. Entropy-regularization (Cuturi, 2013) has emerged as a solution to the computational burden of optimal transport. The entropy-regularized discrete optimal transport problem is defined as follows:

$$(\mathcal{D}_{\mathcal{MK}}^{\varepsilon}) \quad \min_{\gamma \in U(a,b)} \langle \gamma, C \rangle_F - \varepsilon \mathcal{H}(\gamma) \tag{6}$$

where $\mathcal{H}(\gamma) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}(\log(\gamma_{ij}) - 1)$ is the entropy of γ. This regularization allows a faster computation of the optimal transport plan (Peyré et al., 2019) in $\mathcal{O}(r^2/\varepsilon^3)$ (Altschuler et al., 2017) via the iterative procedure of Sinkhorn algorithm (Knight, 2008).

Hierarchical optimal transport is an attractive formulation that offers an efficient way to induce structural information directly into the transportation process (Schmitzer and Schnörr, 2013). The main underlying idea behind this formulation is to organize the data in X and Y into structures (e.g., classes or clusters), this hierarchical organization allows to look at both X and Y as a collection of structures. To compute the hierarchical optimal transport plan between these two collections, the cost function can no longer be evaluated using a distance that quantitatively defines the closeness between data, such as the Euclidean distance, we must therefore employ another metric able to measure the discrepancy between structures. Since each structure can be represented by a discrete measure, the Wasserstein distance is an evident choice. Obviously, computing the Wasserstein distance between each pair of structures requires solving a prior optimal transport problem between samples of the two structures. Therefore, if X and Y are composed of h and l structures respectively, then, the Wasserstein cost matrix would require a prior computation of $h \times l$ optimal transport problems, before solving the final optimal transport problem between classes and clusters, hence the hierarchy.

More formally, let $\mathcal{X}$ be a Polish metric space endowed with a distance d and $\mathcal{P}(\mathcal{X})$ be the space of Borel probability measures on $\mathcal{X}$ equipped with the Wasserstein distance W_p according to (3). Since $\mathcal{X}$ is a Polish metric space, then $\mathcal{P}(\mathcal{X})$ is also a Polish metric space (Parthasarathy, 2005). By a recursion of concepts, $\mathcal{P}(\mathcal{P}(\mathcal{X}))$ the space of Borel probability measures on $\mathcal{P}(\mathcal{X})$ is a Polish metric space, and will be equipped then with the Wasserstein metric that we note HW_p, induced this time by the Wasserstein distance W_p which acts as the ground metric on $\mathcal{P}(\mathcal{X})$. Let $\theta = \{\mu_1, \ldots, \mu_h\} \subset \mathcal{P}(\mathcal{X})$ and $\vartheta = \{\nu_1, \ldots, \nu_l\} \subset \mathcal{P}(\mathcal{X})$ be two sets of probability measures over $\mathcal{P}(\mathcal{X})$ (each probability measure represents a structure). The empirical distributions of θ and ϑ can be expressed respectively by $\phi, \varphi \in \mathcal{P}(\mathcal{P}(\mathcal{X}))$ as $\phi = \sum_{i=1}^{h}\alpha_i\delta_{\mu_i}$ and $\varphi = \sum_{j=1}^{l}\beta_j\delta_{\nu_j}$, where $\alpha = (\alpha_1, \ldots, \alpha_h)$ and $\beta = (\beta_1, \ldots, \beta_l)$ are vectors in the probability simplex $\sum_h$ and $\sum_l$ respectively (ϕ and φ represent the two collections of structures). The hierarchical optimal transport problem between ϕ and φ is then:

$$(\mathcal{HOT}) \quad \min_{\Gamma \in U(\alpha,\beta)} \langle \Gamma, \mathcal{W} \rangle_F \tag{7}$$

where the matrix $\mathcal{W} = (\mathrm{W}_p(\mu_i, \nu_j))_{\substack{1 \le i \le h \\ 1 \le j \le l}} \in \mathcal{M}_{h \times l}(\mathbb{R}^+)$ stands for the Wasserstein cost matrix and $U(\alpha, \beta)$ represents the new transportation polytope, $U(\alpha, \beta) = \{\Gamma \in \mathcal{M}_{h \times l}(\mathbb{R}^+) | \Gamma \mathbb{1}_l = \alpha \text{ and } \Gamma^{\mathbf{T}} \mathbb{1}_h = \beta\}$. More intuitive insights are provided in Fig. 5.

4 HOT-DA: hierarchical optimal transport for unsupervised domain adaptation

In this section, we introduce the proposed HOT-DA approach, which consists of three phases, the first one aims to learn hidden structures in the unlabeled target domain using Wasserstein barycenter, which we prove can be equivalent to spectral clustering, the second phase focuses on finding a one-to-one matching between structures of the two domains through the hierarchical optimal transport formulation, and the third phase involves transporting samples of each source structure to its corresponding target structure via the barycentric mapping.

4.1 Learning unlabeled target structures through Wasserstein-Spectral clustering

Samples in the source domain $S = \{(x_i, y_i)\}_{i=1}^{n}$ can be grouped into structures according to their class labels, but, data in the target domain $T = \{x_j\}_{j=1}^{m}$ are not labeled to allow us to identify directly such structures. Removing this obstacle cannot be accomplished without using some additional assumptions. In fact, to exploit efficiently the unlabeled data in the target domain, the most plausible assumption stems from the structural hypothesis based on clustering, where it is assumed that the data belonging to the same cluster are more likely to share the same label. This assumption constitutes the core nucleus for the first phase of our approach, which aims to prove that spectral clustering can be cast as a problem of learning probability measures with respect to Wasserstein barycenter. Our proof is based on three key ingredients: the equivalence between the search for a 2-Wasserstein barycenter of the empirical distribution that represents unlabeled data and k-means clustering, the analogy between traditional k-means and kernel k-means and finally the connection between kernel k-means and spectral clustering. We derive from this result a novel algorithm able to learn efficiently hidden structures of arbitrary shapes in the unlabeled target domain.

Firstly, given m unlabeled instances $\{x_1, \dots, x_m\} \subset \mathcal{X}$, k-means clustering (MacQueen et al., 1976) aims to partition the m samples into k clusters $\Pi_k = \{\pi_1, \dots, \pi_k\}$ in which each sample belongs to the cluster with the nearest center. This results in a partitioning of the data space into Voronoi cells $(\text{Vor}_q)_{1 \le q \le k}$ generated by the cluster centers $\tilde{C}_k = \{c_1, \dots, c_k\}$. The goal of k-means then is to minimize the mean squared error, and its objective function is defined as:

$$\min_{c_1,\dots,c_k} \frac{1}{m} \sum_{i=1}^{m} \|x_i - c_j\|^2 \tag{8}$$

Let $\hat{\rho}_m = \sum_{i=1}^{m} \frac{1}{m} \delta_{x_i}$ be the empirical distribution of $\{x_1, \dots, x_m\}$. Since $\frac{1}{m} \sum_{i=1}^{m} \|x_i - c_j\|^2 = \mathbb{E}_{x \sim \hat{\rho}_m} \|x - \tilde{C}_k\|^2$, then according to Canas and Rosasco (2012):

$$\frac{1}{m} \sum_{i=1}^{m} \|x_i - c_j\|^2 = \text{W}_2^2(\hat{\rho}_m, \pi_{\tilde{C}_k} \# \hat{\rho}_m) \tag{9}$$

where $\pi_{\tilde{C}_k} : \mathcal{X} \to \tilde{C}_k$ is the projection function mapping each $x \in \text{Vor}_q \subset \mathcal{X}$ to c_q. Since k-means minimizes (9), it also finds the measure that is closest to $\hat{\rho}_m$ among those with support of size k (Pollard, 1982). Which proves the equivalence between k-means and searching for a 2-Wasserstein barycenter of $\hat{\rho}_m$ in $\mathcal{P}_k(\mathcal{X})$, i.e., a minimizer in $\mathcal{P}_k(\mathcal{X})$ of:

$$f(\kappa) = \text{W}_2^2(\hat{\rho}_m, \kappa) \tag{10}$$

Secondly, k-means suffers from a major drawback, namely that it cannot separate clusters that are nonlinearly separable in the input space. Kernel k-means (Schölkopf et al., 1998) can overcome this limitation by mapping the input data in $\mathcal{X}$ to a high-dimensional reproducing kernel Hilbert space $\mathcal{H}$ by a nonlinear mapping $\psi : \mathcal{X} \to \mathcal{H}$, then the traditional k-means is applied on the high-dimensional mappings $\{\psi(x_1), \dots, \psi(x_m)\}$ to obtain a nonlinear partition. Thus, the objective function of kernel k-means can be expressed analogously to that of traditional k-means in (8):

$$\min_{c_1,\dots,c_k} \frac{1}{m} \sum_{i=1}^{m} \|\psi(x_i) - c_j\|^2 \tag{11}$$

Usually, the nonlinear mapping $\psi(x_i)$ cannot be explicitly computed, instead, the inner product of any two mappings $\psi(x_i)^T \psi(x_j)$ can be computed by a kernel function $\mathcal{K}$. Hence,

the whole data set in the high-dimensional space can be represented by a kernel matrix $K \in \mathcal{M}_m(\mathbb{R}^+)$, where each entry is defined as: $K_{i,j} = \mathcal{K}(x_i, x_j) = \psi(x_i)^T \psi(x_j)$.

Thirdly, according to Zha et al. (2001), the objective function of kernel k-means in (11) can be transformed to the following spectral relaxed maximization problem:

$$\max_{Y^T Y = I_k, Y \geq 0} \text{trace}(Y^T K Y) \tag{12}$$

On the other hand, spectral clustering has emerged as a robust approach for data clustering (Shi and Malik, 2000; Ng et al., 2002). Here we focus on the normalized cut for k-way clustering objective function (Gu et al., 2001; Stella and Shi, 2003). Let $G = (V, E, \tilde{K})$ be a weighted graph, where $V = \{x_1, \ldots, x_m\}$ is the vertex set, E the edge set, and $\tilde{K}$ the affinity matrix defined by a kernel $\tilde{\mathcal{K}}$. The k-way normalized cut spectral clustering aims to find a disjoint partition $\{V_1, \ldots, V_k\}$ of the vertex set V, such that:

$$\min_{V_1, \ldots, V_k} \sum_{l=1}^{k} \text{linkratio}(V_l, \overline{V_l}) \tag{13}$$

where $\text{linkratio}(V_l, \overline{V_l}) = \frac{\text{links}(V_l, \overline{V_l})}{\text{degree}(V_l)} = \frac{\sum_{i \in V_l} \sum_{j \in \overline{V_l}} \tilde{K}_{ij}}{\sum_{i \in V_l} \sum_{j \in V} \tilde{K}_{ij}}$.

Following (Dhillon et al., 2004; Ding et al., 2005), the minimization in (13) can be casted as:

$$\max_{Z^T Z = I_k, Z \geq 0} \text{trace}(Z^T \tilde{D}^{-1/2} \tilde{K} \tilde{D}^{-1/2} Z) \tag{14}$$

where $\tilde{D}$ is the degree matrix of the graph G. Thus, the maximization problem in (14) is identical to the spectral relaxed maximization of kernel k-means clustering in (12) when equipped with the kernel matrix $K = \tilde{D}^{-1/2} \tilde{K} \tilde{D}^{-1/2}$.

According to the three-dimensional analysis above, we can now give the main result in the first phase of our method:

Theorem 1 *Spectral clustering using an affinity matrix $\tilde{K}$ is equivalent to the search for a 2-Wasserstein barycenter of $\hat{\varrho}_m = \sum_{i=1}^{m} \frac{1}{m} \delta_{\xi(x_i)}$ in the space of probability measures with support of size k, where ξ is a nonlinear mapping corresponding to the kernel matrix $K = \tilde{D}^{-1/2} \tilde{K} \tilde{D}^{-1/2}$ and $\tilde{D}$ is the degree matrix associated to $\tilde{K}$.*

In the sequel, we will refer to the search for a 2-Wasserstein barycenter of $\hat{\varrho}_m$ as Wasserstein-Spectral clustering, and we will use it to learn k hidden structures in the unlabeled target domain T.

The theoretical result in Theorem 1 is confirmed by experiments, this is illustrated in Fig. 4, where we show that Wasserstein-Spectral clustering performs identically to the traditional spectral clustering and that both are effective at separating clusters that are nonlinearly separable, whereas k-means fails to separate data with non-globular structures.

Complexity analysis: Wasserstein-Spectral clustering offers an alternative to the popular spectral clustering algorithm of Ng et al. (2002) that has limited applicability to large-scale problems due to its prohibitive running time that might be cubic $\mathcal{O}(m^3)$ on the size m of the input dataset (Yan et al., 2009; Tsironis et al., 2013). In fact, there are fast and efficient algorithms to perform Wasserstein-Spectral clustering as Cuturi and Doucet (2014), Kroshnin et al. (2019) which is based on accelerated gradient descent with complexity

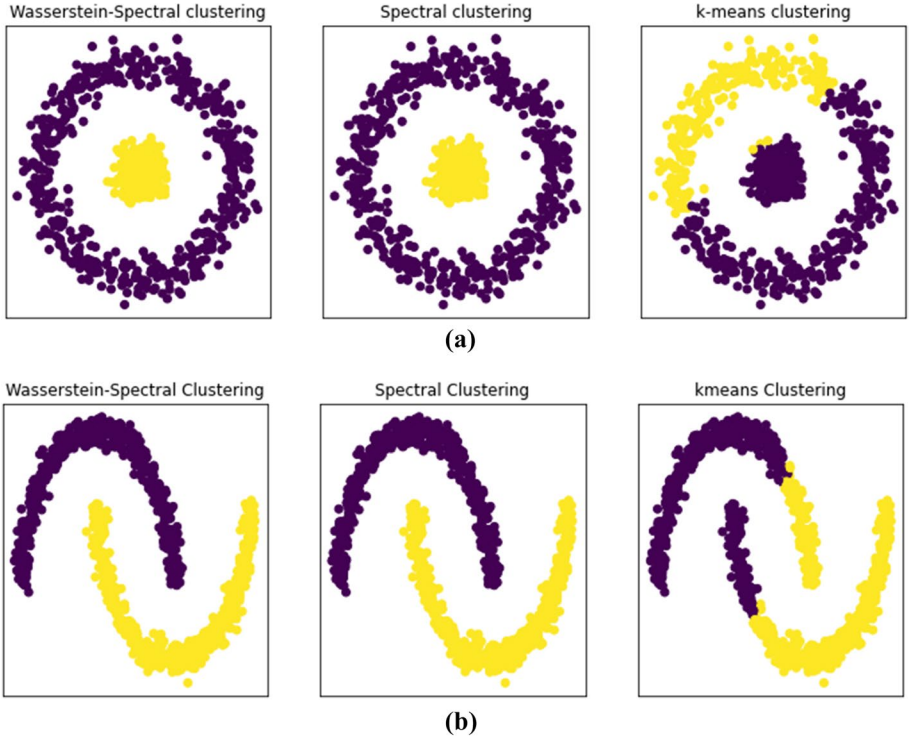


Fig. 4 (**a**) Comparison of Wasserstein-Spectral clustering, spectral clustering, and *k*-means on Two-Circles dataset. (**b**) As for (**a**) but on Moons dataset

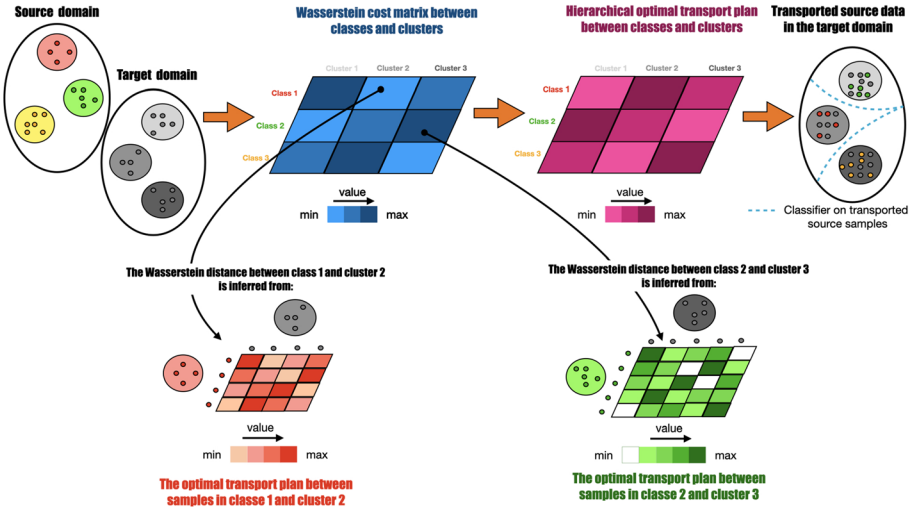


Fig. 5 Wasserstein-Spectral clustering is used to learn hidden structures in the target domain as a seminal step before performing hierarchical optimal transport to align the source and target domains. The optimal plan of this hierarchical transport (in purple) is calculated from the Wasserstein cost matrix (in blue) that measures the distance between the source classes and the target clusters. The distance between each pair of structures is computed through the optimal transport plan of their points (e.g., orange and green) (Color figure online)

proportional to m^2/ε and Altschuler and Boix-Adsera (2021) which can be computed in polynomial time in fixed dimension d. Furthermore, when the barycenter is restricted to measures with support of size k, the recent work of Izzo et al. (2021) shows that randomized dimensionality reduction can be used to map the problem to a space of dimension $\mathcal{O}(\log(k))$ independent of d and that any solution found in the reduced dimension will have its cost preserved up to arbitrary small error in the original space. The algorithmic application of this statement is that one can take any approximation algorithm or heuristic for computing Wasserstein barycenter and combine it with dimensionality reduction to cope with the curse of dimensionality burden of Wasserstein barycenter.

It is noteworthy that the computation of Wasserstein barycenter is an increasingly popular problem in the machine learning and statistics communities and our algorithm can benefit from this renewed interest to reach more faster running time.

4.2 Matching source and target structures through hierarchical optimal transport

Optimal transport offers a well-founded geometric way for comparing probability measures in a Lagrangian framework, and for inferring a matching between them as an inherent part of its computation. Its hierarchical formulation has inherited all these properties with the extra benefit of inducing structural information directly without the need to add any regularized term for this purpose, as well as the capability to split a sophisticated optimization surface into simpler ones that are less subject to local minima, and the ability to benefit from the entropy-regularization. Hence the key insight behind its use in the second phase of our method.

To use an appropriate formulation for hierarchical optimal transport, samples in the source domain $S = \{(x_i, y_i)\}_{i=1}^{n}$ must be partitioning according to their class labels $y_i \in \mathcal{Y} = \{c_1, \ldots, c_k\}$ into k classes $\{C_1, \ldots, C_k\}$. The empirical distributions of these structures can be expressed using discrete measures $\{\mu_1, \ldots, \mu_k\} \subset \mathcal{P}(\mathcal{X})$ as follows:

$$\mu_h = \sum_{i=1/x_i \in C_h}^{n} a_i \delta_{x_i}, \quad \forall h \in \{1, \ldots, k\} \tag{15}$$

Similarly, samples in the target domain $T = \{x_j\}_{j=1}^{m}$ are grouped in k clusters $\{Cl_1, \ldots, Cl_k\}$ using Wasserstein-Spectral clustering in the first phase. The empirical distributions of these structures can be expressed using discrete measures $\{\nu_1, \ldots, \nu_k\} \subset \mathcal{P}(\mathcal{X})$ in the following way:

$$\nu_l = \sum_{j=1/x_j \in Cl_l}^{m} b_j \delta_{x_j}, \quad \forall l \in \{1, \ldots, k\} \tag{16}$$

Under the assumption that S and T are two sets of independent and identically distributed samples, the weights of all instances in each structure are naturally set to be equal:

$$a_i = \frac{1}{|C_h|} \quad \text{and} \quad b_j = \frac{1}{|Cl_l|}, \quad \forall h, l \in \{1, \ldots, k\}$$

The set S of labeled source samples and the set T of unlabeled target samples can be seen in a hierarchical paradigm as a collection of classes and clusters. Thus, the distribution of S and T can be expressed respectively as a measure of measures ϕ and φ in $\mathcal{P}(\mathcal{P}(\mathcal{X}))$ as follows:

$$\phi = \sum_{h=1}^{k} \alpha_h \delta_{\mu_h} \quad \text{and} \quad \varphi = \sum_{l=1}^{k} \beta_l \delta_{\nu_l} \tag{17}$$

where $\alpha = (\alpha_1, \ldots, \alpha_k)$ and $\beta = (\beta_1, \ldots, \beta_k)$ are vectors in the probability simplex $\sum_k$. The weights α_h and β_l are set to be equal to deal with the problem of structure imbalance, in the following way:

$$\alpha_h = \frac{1}{k} \quad \text{and} \quad \beta_l = \frac{1}{k}, \quad \forall h, l \in \{1, \ldots, k\}$$

To learn the correspondences between classes and clusters, we formulate an entropy-regularized hierarchical optimal transport problem between ϕ and φ in the following way:

$$(\mathcal{HOT} \diagdown \mathcal{DA}) \quad \min_{\Gamma \in U(\alpha,\beta)} \langle \Gamma, \mathcal{W} \rangle_F - \varepsilon \mathcal{H}(\Gamma) \tag{18}$$

where $U(\alpha, \beta) = \{\Gamma \in \mathcal{M}_k(\mathbb{R}^+) | \Gamma \mathbb{1}_k = \alpha \text{ and } \Gamma^{\mathbf{T}} \mathbb{1}_k = \beta\}$ represents the transportation polytope and $\mathcal{W} = (\mathcal{W}_{h,l})_{1 \leq h,l \leq k} \in \mathcal{M}_k(\mathbb{R}^+)$ stands for the Wasserstein cost matrix, whose each matrix-entry $\mathcal{W}_{h,l}$ is defined as the *2*-Wasserstein distance between the measures μ_h and ν_l:

$$\mathcal{W}_{h,l}^2 = \mathrm{W}_2^2(\mu_h, \nu_l) = \langle \gamma_{h,l}^{*,\varepsilon\prime}, \mathcal{C}_{h,l} \rangle_F \tag{19}$$

where $\mathcal{C}_{h,l}$ is the cost matrix of pairwise squared-Euclidean distances between elements of C_h and Cl_l, and $\gamma_{h,l}^{*,\varepsilon\prime}$ is the regularized optimal transport plan between μ_h and ν_l.

The optimal transport plan Γ_ε^* in (18) can be interpreted as a soft multivalued matching between ϕ and φ as it provides the degree of association between classes $\{C_1, \ldots, C_k\}$ in the source domain S and clusters $\{Cl_1, \ldots, Cl_k\}$ in the target domain T. Then, the one-to-one matching relationship ($\widehat{=}$) between each class C_h and its corresponding cluster Cl_l can be inferred by hard assignment from Γ_ε^*, in the following way:

$$C_h \widehat{=} Cl_l | l = \underset{j=1,\ldots,k}{\operatorname{argmax}} \Gamma_\varepsilon^*(h, j), \quad \forall h \in \{1, \ldots, k\} \tag{20}$$

4.3 Transporting source to target structures through the barycentric mapping

Besides being a means of comparison and matching, optimal transport has the asset of performing thanks to its intrinsic quiddity of transport an alignment between source and target structures. Hence the main underlying idea of this phase.

Once the correspondence between source and target structures has been determined according to the one-to-one matching relationship ($\widehat{=}$) in (20), the source samples in each class C_h have to be transported to the target samples in the corresponding cluster Cl_l. This transportation can be handily expressed for each instance x_i in C_h with respect to the instances in Cl_l as the following barycentric mapping (Reich, 2013; Ferradans et al., 2014; Courty et al., 2016):

$$\widetilde{x_i} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \sum_{j=1/x_j \in Cl_l}^{m} \gamma_{h,l}^{*,\varepsilon\prime}(i,j)\|x - x_j\|^2 \tag{21}$$

where $\widetilde{x_i}$ is the image of x_i in the region occupied by Cl_l on the target domain, and $\gamma_{h,l}^{*,\varepsilon\prime}$ is the optimal transport plan between μ_h and ν_l already computed in (19). The barycentric mapping can be formulated for each class C_h as follows:

$$\widetilde{C_h} = diag(\gamma_{h,l}^{*,\varepsilon\prime} \mathbb{1}_{|Cl_l|})^{-1} \gamma_{h,l}^{*,\varepsilon\prime} Cl_l, \quad \forall h \in \{1, \ldots, k\} \tag{22}$$

While samples in C_h and Cl_l are drawn i.i.d. from μ_h and ν_l, then this mapping can be casted as a linear expression:

$$\widetilde{C_h} = |C_h| \gamma_{h,l}^{*,\varepsilon\prime} Cl_l, \quad \forall h \in \{1, \ldots, k\} \tag{23}$$

After the alignment of each class C_h with its corresponding cluster Cl_l has been done as suggested in (23), a classifier η can be learned on the transported labeled source data $\widetilde{S} = \cup_{q=1}^{k} \widetilde{C_q}$ and evaluated on the unlabeled target data T.

The proposed HOT-DA approach is formally summarized in Algorithm 1:

Algorithm 1 HOT-DA

Input : $S = \{(x_i, y_i)\}_{i=1}^{n}, T = \{x_j\}_{j=1}^{m}$

Parameter: $\varepsilon, \varepsilon\prime$

1: Form $\mu_h, \nu_l \quad \forall h, l \in \{1, ..., k\}$ (15)(16)
2: Form ϕ, φ (17)
3: Solve the HOT-DA problem between ϕ and φ (18)
4: Get the one-to-one matching between structures (20)
5: Transport the source structures to the target ones to get $\widetilde{S}$ (23)
6: Train a classifier η on $\widetilde{S}$ and evaluate it on T
7: **return** $\{y_j\}_{j=1}^{m}$

5 Experimental results

In this section, we evaluate our method on a toy dataset and three challenging real-world visual adaptation problems.[1]

5.1 Inter-twinning moons dataset

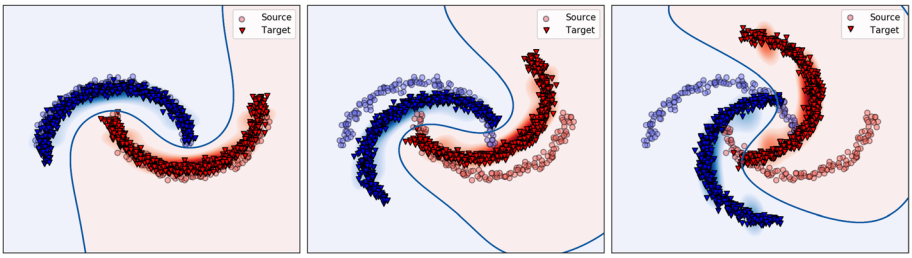
In the first experiment, we carry on moons dataset, the source domain is the classical binary two inter-twinning moons centered at the origin (0,0) and composed of 300 instances, where each class is associated to one moon of 150 samples. We consider 7 different target domains by rotating anticlockwise the source domain around its center according to 7 angles.

[1] We make our code and the used datasets publicly available at: https://github.com/MouradElHamri/HOT-DA.

Table 1 Average accuracy over moons dataset for 7 rotation angles

Angle (°)	10°	20°	30°	40°	50°	70°	90°
SVM	**1**	0.896	0.760	0.688	0.600	0.266	0.172
PBDA	**1**	0.906	0.897	0.775	0.588	0.374	0.313
OT-GL	**1**	**1**	**1**	0.987	0.804	0.622	0.492
JDOT	0.989	0.955	0.906	0.865	0.815	0.705	0.600
HiWA	0.575	0.579	0.514	0.579	0.579	0.552	0.399
MADAOT	0.995	0.993	0.996	0.996	0.989	0.770	0.641
HOT-DA	**1**	**1**	**1**	**1**	**1**	**1**	**0.997**

Bolded numbers correspond to the best performance



(a) rotation =10° **(b)** rotation =40° **(c)** rotation =90°

Fig. 6 Illustration of the decision boundary of HOT-DA over moons problem for increasing rotation angles

Naturally, the greater is the angle, the harder is the adaptation. The experiments were run by setting $\varepsilon = \varepsilon, = 0.1$, and an SVM with a Gaussian kernel as classifier to cope with the non-linearity of this dataset. The width parameter of the SVM was chosen as $\sigma = \frac{1}{2\mathbb{V}}$, where $\mathbb{V}$ is the variance of the transported source samples. Our algorithm is compared to an SVM classifier with a Gaussian kernel trained on the source domain (without adaptation), PBDA (Germain et al., 2013) and four optimal transport based domain adaptation methods, OT-GL (Courty et al., 2016), JDOT (Courty et al., 2017), HiWA (Lee et al., 2019) and MADAOT (Dhouib et al., 2020), with the hyperparameter ranges suggested in the respective articles. To assess the generalization ability of the compared methods, they are tested on an independent set of 1000 instances that follow the same distribution as the target domain. The experiments are conducted 10 times, and the average accuracy is considered as a comparison criterion. The results are presented in Table 1 and the decision boundary of HOT-DA is illustrated in Fig. 6.

We remark that all the considered algorithms based on optimal transport (except for HiWa) manage to achieve an almost perfect score on the angles from 10° to 40°, which is rational, as for these small angles the adaptation problem remains quite easy. However, the SVM without adaptation has experienced a decline of almost one-third of its accuracy from 30°. This proves that moons dataset presents a difficult adaptation problem that goes beyond the generalization ability of standard supervised learning models. For the strongest deformation, from 50° and up to 90°, the proposed method HOT-DA, always provides an almost perfect score, while a big deterioration in the performance

Table 2 Description of the visual adaptation datasets

Dataset	Domains	#Samples	#Features	#Classes	Abbr.
Digits	USPS	1800	256	10	U
	MNIST	2000	256	10	M
Office-Caltech	Caltech	1123	4096	10	C
	Amazon	958	4096	10	A
	Webcam	295	4096	10	W
	DSLR	157	4096	10	D
Office-Home	Art	2427	2048	65	Ar
	Clipart	4365	2048	65	Cl
	Product	4439	2048	65	Pr
	Real-World	4357	2048	65	Rw

of PBDA and considerable deterioration in the performance of OT-GL and JDOT from 50° was observed, for MADAOT, a significant deterioration of performances starts from 70°. In short, structures leveraged by HOT-DA are highlighted by eliminating the increasing difficulty of this adaptation task, the constancy of the excellent performances of our approach speaks for itself, while the poor performances of HiWa, which is a multimodal distribution alignment method that seeks to jointly learn the alignment and the structure-correspondences is rather surprising, considering that this approach also relies on hierarchical optimal transport.

5.2 Visual adaptation datasets

We now evaluate our method on three challenging visual adaptation datasets. We start by presenting the details of these benchmark datasets, the experimental protocol, the hyperparameter tuning and finish by providing and discussing the obtained results.

Datasets: We consider three visual adaptation datasets: Digits (Hull, 1994; LeCun, 1998), Office-Caltech (Fei-Fei et al., 2004; Saenko et al., 2010) and Office-Home (Venkateswara et al., 2017). A detailed description of each dataset is given in Table 2.

Experimental protocol: For the problem of Digits recognition, 2000 and 1800 images are randomly selected respectively from the original MNIST and USPS datasets. Then, the selected MNIST images are resized to the same 16×16 resolution as USPS ones. For the second visual adaptation problem, Office-Caltech dataset is used, where we randomly sampled a collection of 20 images per class from each domain, except for DSLR where only 8 images per class are selected. To represent these images, 4096 DeCaf6 features are used (Donahue et al., 2014). For the last problem, the more complex Office-Home dataset (Venkateswara et al., 2017) is employed. This dataset contains 15588 images from four visually very different domains: Artistic images, Clip Art, Product images, and Real-world images. For this problem, ResNet-50 was used to extract 2048 features (He et al., 2016).

As a classifier for our approach, we use 1-Nearest Neighbor classifier (1NN) on the three visual adaptation datasets, which has the advantage of being parameter free.

For the problem of Digits recognition, the comparison is conducted using 1NN classifier (without adaptation) and five domain adaptation methods, SA (Fernando et al., 2013) with a linear SVM, JDA (Long et al., 2013) with 1NN classifier, SCA (Ghifary et al., 2016) with 1NN classifier, OT-GL with 1NN classifier (Courty et al., 2016) and JDOT with a linear SVM (Courty et al., 2017). Concerning Office-Caltech dataset, the comparison is

Table 3 Accuracy on digits dataset

Task	1NN	JDA	SA	SCA	OT-GL	JDOT	**HOT-DA**
M → U	58.33	60.09	67.71	65.10	<u>69.96</u>	64.00	**76.39**
U → M	39.00	54.52	49.85	48.00	<u>57.85</u>	56.00	**63.20**
average	48.66	57.30	58.73	56.55	<u>63.90</u>	60.00	**69.79**

Bolded numbers correspond to the best performance and underlined numbers to the second best performance

Table 4 Accuracy on Office-Caltech dataset (Decaf6 features)

Task	1NN	JDA	SA	SCA	OT-GL	JDOT	DeepJDOT	**HOT-DA**
A → C	22.25	81.28	79.20	78.80	<u>85.51</u>	85.22	**87.40**	80.00
A → D	20.38	86.25	83.80	85.40	85.00	87.90	<u>88.50</u>	**92.53**
A → W	23.51	<u>88.33</u>	74.60	75.90	83.05	84.75	86.70	**96.74**
C → A	20.54	88.04	89.30	89.50	92.08	91.54	**92.30**	<u>92.19</u>
C → D	19.62	84.12	74.40	87.90	87.25	89.91	<u>92.00</u>	**96.27**
C → W	18.94	79.60	88.50	85.40	84.17	<u>88.81</u>	85.30	**95.11**
D → A	27.10	91.32	79.00	90.00	**92.31**	88.10	<u>91.50</u>	91.33
D → C	23.97	81.13	**92.25**	78.10	84.11	84.33	<u>85.30</u>	78.48
D → W	51.26	97.48	79.20	<u>98.60</u>	96.29	96.61	**98.70**	96.33
W → A	23.19	90.19	55.00	86.10	90.62	<u>90.71</u>	86.60	**91.86**
W → C	19.29	81.97	**99.60**	74.80	81.45	82.64	<u>84.70</u>	78.20
W → D	53.62	<u>98.88</u>	81.65	**100.00**	96.25	98.09	98.70	94.61
average	28.47	86.72	81.65	85.90	88.18	89.05	<u>89.80</u>	**90.30**

Bolded numbers correspond to the best performance and underlined numbers to the second best performance

performed with the same competitors as for Digits in addition to DeepJDOT (Damodaran et al., 2018). Regarding the more voluminous and challenging Office-Home dataset, the choice is made to conduct the comparison with five deep learning approaches to prove the scalability of our method, and its capability to compete with deep learning models. The competitors are: ResNet-50 (without adaptation), DAN (Long et al., 2015), DANN (Ganin et al., 2016), JAN (Long et al., 2017) and DeepJDOT (Damodaran et al., 2018).

Hyper-parameter tuning: For the problem of Digits recognition, the experiments were performed by setting $\varepsilon = \varepsilon, = 0.1$. For Office-Caltech dataset, each target domain is equitably splited into a validation and test set. The validation set is used to select the best hyper-parameters $\varepsilon, \varepsilon,$ in the range of $\{1, \ldots, 100\}$. The accuracy is then evaluated on the test set, with the chosen hyper-parameters. The experimentation is performed 10 times, and the mean accuracy in % is reported as in Courty et al. (2016). For Office-Home dataset, all labeled source samples and unlabeled target samples are used, and the average classification accuracy in % is computed based on three random experiments as in Ganin and Lempitsky (2015). The best hyper-parameters $\varepsilon, \varepsilon,$ are selected in the range of $\{1, \ldots, 100\}$.

Results: The results of our experiments are reported in Tables 3, 4, and 5. For each task, we use bold and underlined fonts to indicate the best and second best results respectively.

From Table 3, we can see that the proposed approach HOT-DA significantly outperforms the other domain adaptation methods on both tasks of Digits recognition problem.

Table 5 Accuracy on Office-Home dataset (ResNet-50 features)

Task	ResNet-50	DAN	DANN	JAN	DeepJDOT	**HOT-DA**
Ar → Cl	34.9	43.6	45.6	45.9	**50.7**	48.0
Ar → Pr	50.0	57.0	59.3	61.2	68.6	**69.0**
Ar → Rw	58.0	67.9	70.1	68.9	74.4	**75.3**
Cl → Ar	37.4	45.8	47.0	50.4	59.9	**61.7**
Cl → Pr	41.9	56.5	58.5	59.7	**65.8**	63.2
Cl → Rw	46.2	60.4	60.9	61.0	**68.1**	67.4
Pr → Ar	38.5	44.0	46.1	45.8	**55.2**	54.1
Pr → Cl	31.2	43.6	43.7	43.4	**46.3**	39.7
Pr → Rw	60.4	67.7	68.5	70.3	73.8	**75.3**
Rw → Ar	53.9	63.1	63.2	63.9	66.0	**67.6**
Rw → Cl	41.2	51.5	51.8	52.4	**54.9**	47.9
Rw → Pr	59.9	74.3	76.8	76.8	78.3	**78.5**
average	46.1	56.3	57.6	58.3	**63.5**	62.4

Bolded numbers correspond to the best performance and underlined numbers to the second best performance

Table 4 shows that HOT-DA surpasses the other competitors on 5 out of 12 tasks in Office-Caltech dataset, and has the second best accuracy on another task. Tables 3 and 4 also present the average results of each algorithm, where we observe a slight advance in favor of our method compared to competitors, notably JDOT and DeepJDOT. Therefore, we attribute this gain to the effectiveness of our Wasserstein-Spectral clustering that succeeds in learning hidden structures in the target domain even if they do not have globular shapes, which is the case of these two challenging visual adaptation datasets. Furthermore, the hierarchical formulation incorporates efficiently these structures, which allows to preserve compact classes during the transportation and limits the mass splitting across different target structures. However, we see that DeepJDOT significantly outperforms HOT-DA in the three tasks where Caltech (C) is the target domain, this is explained by the difficulty we encountered to produce clusters similar to the unknown real classes in this domain.

The experimental results on Office-Home dataset are shown in Table 5. We observe that HOT-DA outperforms the other methods on 6 out of 12 tasks, while DeepJDOT performs better in the remaining 6 tasks. DeepJDOT is in the second place 6 times compared to 3 times for HOT-DA, which experienced a drop in performance in the 3 tasks where Clipart is the target domain. This behavior led to a slight difference in their average accuracy on Office-Home dataset in favor of DeepJDOT. This is rather surprising considering that the competitors rely on neural networks to learn the final classifier and these latter are expected to have higher discriminative power than the 1-Nearest Neighbor classifier used in our approach. Consequently, we attribute this competitiveness to the efficiency of our hierarchical optimal transport formulation that manages to better align the two distributions, and that can be seen as an "implicit regularized" optimal transport. This implicit regularization heavily relies on "a priori knowledge" (clustering), which leads to the injection of structural information directly into the transport problem.

Globally, the mean accuracy of HOT-DA is 0.5% higher than DeepJDOT on Office-Caltech. In parallel, DeepJDOT shows an improvement of 1.1% compared to our method on Office-Home. Roughly speaking, the set of experiments shows a good behavior with respect to state-of-the-art methods, especially JDOT and DeepJDOT, which however

manage to outperform our algorithm on several tasks. This competitive behavior is, we believe, due to the commonality between JDOT and DeepJDOT on the one hand and HOT-DA on the other hand. The former methods design a simultaneous optimization problem to find the coupling between the joint distribution of the source and target domains and the labeling function that solves the transfer problem. While the second method tries to address the same task sequentially by first finding the target structures, which is equivalent to performing a pseudo-labeling in the target domain, before aligning each source structure with its corresponding target structure, which can be seen as an alignment of the joint distributions.

5.3 Relevance of Wasserstein-Spectral clustering to HOT-DA

The first step of HOT-DA is not directly integrated into the domain adaptation process, and it is questionable whether other well-known clustering algorithms such as k-means (MacQueen et al., 1976), DBSCAN (Ester et al., 1996) or HDBSCAN (Campello et al., 2013) can be used to learn the target structures instead of Wasserstein-Spectral clustering (W-SC).

k-means suffers from several drawbacks, notably its inability to identify clusters with non-convex shapes, as shown in Fig. 4. This incapacity can significantly reduce the performance of HOT-DA on several unsupervised domain adaptation problems where clusters do not have globular shapes in the target domain. These problems include but are not limited to, the inter-twinning moons dataset.

On the other hand, DBSCAN relies on detecting areas where points are closely packed together (points with many nearby neighbors) and marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN does not require to specify the number of clusters a priori, instead, it requires two parameters: minimum number of neighbors *minpts* and minimum radius *Eps*. Therefore, for clustering high-dimensional data, it becomes very difficult to tune these parameters to get the desired number of clusters, even using heuristic methods (Musdholifah et al., 2013). Which can lead to finding a number of clusters very larger or very smaller than the number of classes k in the source domain, and then to poor adaptation results. Regarding HDBSCAN, which is a conversion of DBSCAN into a hierarchical clustering algorithm, from which a simplified hierarchy composed only of the most significant clusters can be easily extracted. It can find clusters of varying densities, unlike DBSCAN and it performs well on low to medium dimensional data. However, its performance tends to decrease as the dimension increases. In general, the performance of HDBSCAN can see significant decreases already with tens of dimensions (Campello et al., 2020). The unsupervised domain adaptation settings can be beneficial for clustering algorithms that require the number of clusters k to the detriment of DBSCAN and HDBSCAN which do not benefit from this available information, especially for high-dimensional data (e.g., visual domain adaptation datasets using ResNet-50 or DeCaf features) where it becomes quite difficult to tune these parameters to get the desired number of clusters k.

This analysis is the main motivation behind replacing k-means, DBSCAN, or HDBSCAN with spectral clustering which is able to find exactly k clusters, even with non-globular shapes. This choice was reconsidered for complexity reasons as discussed in Sect. 4.1, which led to the establishment of an equivalent algorithm: Wasserstein-Spectral clustering, which furthermore allows unifying the different steps of our algorithm under the aegis of optimal transport.

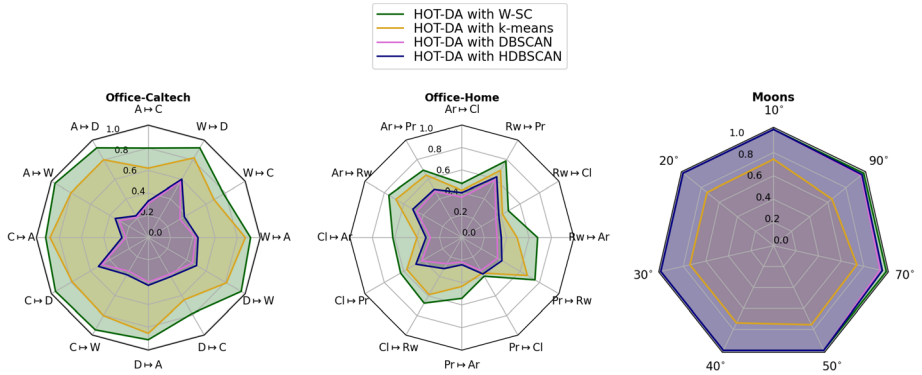


Fig. 7 Kiviat's accuracy diagram for the four variants of HOT-DA on Office-Caltech, Office-Home, and Moons datasets. The radar corresponding to the variant based on Wasserstein-Spectral clustering dominates the other radars on the three datasets

To confirm the insights above, we reproduce the experiments on the following datasets: Moons, Office-Caltech, and Office-Home using four variants of our algorithm, the first one uses Wasserstein-Spectral clustering, the second one uses k-means, the third one is based on DBSCAN and the fourth one is rather based on HDBSCAN. The results of these experiments are given in Fig. 7 using Kiviat diagram.

Figure 7 indicates ostensibly that the radar corresponding to the variant with W-SC encompasses the other radars on the three datasets. On the 7 rotation problems of moons dataset, the variant of HOT-DA based on Wasserstein-Spectral clustering performs slightly better than the other variants based on DBSCAN and HDBSCAN and all manage to make a nearly perfect adaptation. This is due to the ability of Wasserstein-Spectral clustering to capture the structure of the two moons, and the ease of tuning the parameters for DBSCAN and HDBSCAN to find the desired number of clusters in a small dimensional space ($d = 2$). While the variant of HOT-DA based on k-means has much poorer performance due to the inability of k-means to correctly explore the two inter-twinning moons. Regarding Office-Caltech and Office-Home, the high-dimensionality of these datasets ($d = 4096$ for Office-Caltech and $d = 2048$ for Office-Home) has strongly impacted the performance of DBSCAN and HDBSCAN, which fail to find exactly the desired number of clusters ($k = 10$ for Office-Caltech and $k = 65$ for Office-Home), while k-means and Wasserstein-Spectral clustering benefit from this available information to obtain better results, with significant supremacy for this latter.

The above empirical experiments strengthen our choice of Wasserstein-Spectral clustering and clearly demonstrate that it is a well-suited candidate for these unsupervised domain adaptation settings.

5.4 Structure imbalance sensitivity analysis

The problem of structure imbalance where an uneven distribution of samples occurs among a variety of structures can lead to pathological behavior of the mass transportation, by showing favoritism towards majority target structures in spite of minority ones which may receive no mass due to the thresholding performed in (20). Fortunately, the choice made to give the same mass to each structure, allows HOT-DA to avoid this behavior and to achieve

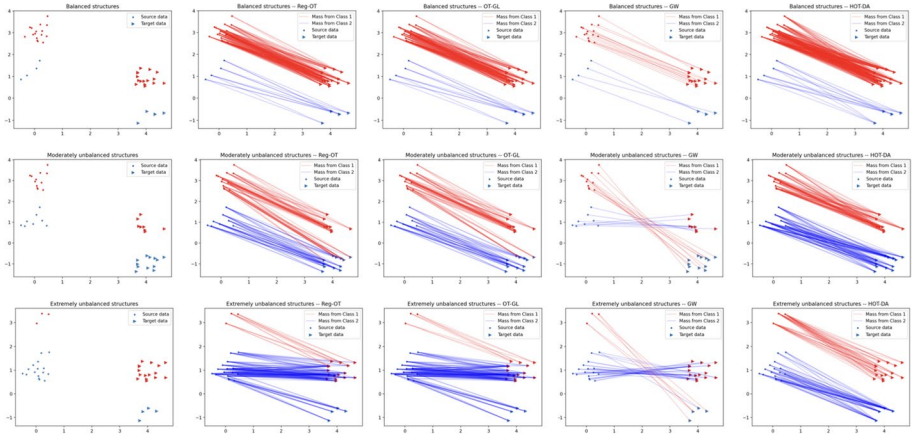


Fig. 8 Behavior of Reg-OT, OT-GL, GW, and, HOT-DA towards the problem of structure imbalance

the right matching between source and target structures. The intuition behind this choice is to consider each structure as an independent entity and to remove the bias induced by its cardinality, which is quite natural since a class in the source domain and its corresponding cluster in the target domain do not necessarily have the same proportion of points.

To evaluate the behavior of HOT-DA with respect to the problem of structure imbalance, an experiment is conducted on a toy dataset composed of two structures in each domain as shown in Fig. 8. The experiment is designed to compare the performance of our proposed approach with Reg-OT (Cuturi, 2013), OT-GL (Courty et al., 2016) and GW (Gromov-Wasserstein is a distance that generalizes the notion of optimal transport to the setting of mm-spaces up to isometries) (Mémoli, 2011; Sturm, 2006), in three scenarios: balanced structures, moderately unbalanced structures and, extremely unbalanced structures.

The first part of the experiment concerning the case of balanced structures shows an ideal behavior of the four methods. The situation begins to change slightly in the second case of moderately unbalanced structures, where Reg-OT and OT-GL make some mistakes because of the extra-mass of the red source structure that has to be sent to the blue target structure, while GW reverses the matching due to this moderate imbalance. However, our approach still achieves an uncontested matching. The third part concerning the most complicated scenario of extremely unbalanced structures, demonstrates a catastrophic deterioration in the results of the three methods Reg-OT, OT-GL, and GW, while our HOT-DA approach continues to provide a flawless result. This proves that HOT-DA is a robust and non-sensitive algorithm to this kind of imbalance, unlike other approaches. It is noteworthy that our model is less sensitive than other optimal transport methods to changes in the value of the entropy regularization parameter thanks to the thresholding carried out by the hard assignment in (20).

6 Conclusions and future perspectives

In this paper, we proposed HOT-DA, a novel approach dealing with unsupervised domain adaptation, by leveraging the ability of hierarchical optimal transport to induce structural information directly into the transportation process. We also proved theoretically the

equivalence between spectral clustering and the problem of learning probability measures through Wasserstein barycenter, this latter was used to derive Wasserstein-Spectral clustering, a new alternative of spectral clustering able to learn hidden structures of arbitrary shapes in the unlabeled target domain, as a seminal step before performing hierarchical optimal transport to align the source and target domains. The proposed approach has been shown to be efficient on both simulated and real-world problems compared to several state-of-the-art methods, in addition to being able to cope with structure imbalance.

Our work can be extended in different directions. From an algorithmic standpoint, we plan to investigate a possible application of the proposed approach to multi-source domain adaptation setting. From a theoretical standpoint, future work will include the development of generalization bounds that take into account the hierarchical organization of source and target samples in structures. These bounds will reflect explicitly both the excess clustering risk in the target domain and which structures must be aligned to lead to a good adaptation.

Author contributions Contributing authors are: Mourad El Hamri, Younès Bennani and Issam Falih.

Funding Not Applicable

Data availibility The used datasets are available at: https://github.com/MouradElHamri/HOT-DA.

Code availability The code is available at: https://github.com/MouradElHamri/HOT-DA.

Declarations

Conflict of interest Not Applicable.

Ethics approval Not Applicable.

Consent to participate Not Applicable.

Consent for publication Not Applicable.

References

Agueh, M., & Carlier, G. (2011). Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis, 43*(2), 904–924.

Altschuler, J., Weed, J., & Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. NIPS'17, pp. 1961-1971.

Altschuler, J. M., & Boix-Adsera, E. (2021). Wasserstein barycenters can be computed in polynomial time in fixed dimension. *Journal of Machine Learning Research, 22,* 44–1.

Alvarez-Melis, D., Jaakkola, T., & Jegelka, S. (2018). Structured optimal transport. In *International conference on artificial intelligence and statistics*, (pp. 1771–1780). PMLR.

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. In: *Advances in neural information processing systems* 19 .

Bertsimas, D., & Tsitsiklis, J. N. (1997). *Introduction to linear optimization* (Vol. 6). Athena Scientific Belmont.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *45th annual meeting of the ACL*.

Campello, R. J., Kröger, P., Sander, J., & Zimek, A. (2020). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(2), e1343.

Campello, R.J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172. Springer.

Canas, G., & Rosasco, L. (2012). Learning probability measures with respect to optimal transport metrics. In *Advances in neural information processing systems*, (Vol. 25).

Chen, Q., Liu, Y., Wang, Z., Wassell, I., & Chetty, K. (2018). Re-weighted adversarial adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 7976–7985).

Courty, N., Flamary, R., Habrard, A., & Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. In *Advances in neural information processing systems*.

Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(9), 1853–1865.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems* 26 .

Cuturi, M., & Doucet, A. (2014). Fast computation of wasserstein barycenters. In *International conference on machine learning*, (pp. 685–693). PMLR.

Damodaran, B.B., Kellenberger, B., Flamary, R., Tuia, D., & Courty, N. (2018). Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, (pp. 447–463).

Dhillon, I., Guan, Y., & Kulis, B. (2004). Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth international conference on Knowledge discovery and data mining*.

Dhouib, S., Redko, I., & Lartizien, C. (2020). Margin-aware adversarial domain adaptation with optimal transport. In *International conference on machine learning*, (pp. 2514–2524). PMLR.

Ding, C., He, X., & Simon, H.D. (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the SIAM international conference on data mining*.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014) Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, (pp. 647–655). PMLR.

Ester, M., Kriegel, H. P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *In kdd, 96,* 226–231.

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, (pp. 178–178). IEEE.

Fernando, B., Habrard, A., Sebban, M., & Tuytelaars, T. (2013). Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE ICCV*, (pp. 2960–2967).

Ferradans, S., Papadakis, N., Peyré, G., & Aujol, J. F. (2014). Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences, 7*(3), 1853–1882.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, (pp. 1180–1189). PMLR.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *JMLR, 17*(1), 2096–2030.

Germain, P., Habrard, A., Laviolette, F., & Morvant, E. (2013). A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *International conference on machine learning*, (pp. 738–746). PMLR.

Ghifary, M., Balduzzi, D., Kleijn, W. B., & Zhang, M. (2016). Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(7), 1414–1430.

Gong, B., Shi, Y., Sha, F., & Grauman, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, (pp. 2066–2073).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *NIPS'14*.

Gu, M., Zha, H., Ding, C., He, X., Simon, H., & Xia, J. (2001). Spectral relaxation models and structure analysis for k-way graph clustering and bi-clustering .

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 770–778).

Ho, N., Nguyen, X., Yurochkin, M., Bui, H.H., Huynh, V., & Phung, D. (2017). Multilevel clustering via wasserstein means. In *International conference on machine learning*, (pp. 1501–1509).

Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 16*(5), 550–554.

Izzo, Z., Silwal, S., & Zhou, S. (2021). Dimensionality reduction for wasserstein barycenter. In *Advances in neural information processing systems* 34 .

Kantorovich, L. V. (1942). On the translocation of masses. *In Doklady Akademii Nauk USSR (NS), 37,* 199–201.

Knight, P.A. (2008). The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications 30*(1) .

Kotsiantis, S.B., Zaharakis, I., & Pintelas, P. et al. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering* 3–24.

Kouw, W. M., & Loog, M. (2019). A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(3), 766–785.

Kroshnin, A., Tupitsa, N., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., & Uribe, C. (2019). On the complexity of approximating wasserstein barycenters. In *International conference on machine learning*, (pp. 3530–3540). PMLR.

LeCun, Y. (1998). The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/ .

Lee, J., Dabagia, M., Dyer, E., & Rozell, C. (2019). Hierarchical optimal transport for multimodal distribution alignment. In *Advances in neural information processing systems*, (Vol. 32).

Li, M., Zhai, Y.M., Luo, Y.W., Ge, P.F., & Ren, C.X. (2020). Enhanced transport distance for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 13936–13944).

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International conference on machine learning*, (pp. 97–105). PMLR.

Long, M., Wang, J., Ding, G., Sun, J., & Yu, P.S. (2013). Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*.

Long, M., Zhu, H., Wang, J., & Jordan, M.I. (2017). Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, (pp. 2208–2217). PMLR.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*.

Mémoli, F. (2011). Gromov-wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics, 11,* 417–487.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris* .

Musdholifah, A., Hashim, S. Z. M., & Zaiton, S. (2013). Cluster analysis on high-dimensional data: A comparison of density-based clustering algorithms. *Australian Journal of Basic and Applied Sciences, 7*(2), 380–389.

Ng, A.Y., Jordan, M.I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, (pp. 849–856).

Pan, S. J., Tsang, I. W., Kwok, J. T., & Yang, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks, 22*(2), 199–210.

Pan, S.J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering 22*(10) .

Parthasarathy, K.R. (2005). *Probability measures on metric spaces*, (Vol 352). American Mathematical Society.

Pele, O., & Werman, M. (2009). Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, (pp. 460–467). IEEE.

Peyré, G., Cuturi, M. et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning 11*(5-6): 355–607 .

Pollard, D. (1982). Quantization and the method of k-means. *IEEE Transactions on Information Theory, 28*(2), 199–205.

Redko, I., Courty, N., Flamary, R., & Tuia, D. (2019). Optimal transport for multi-source domain adaptation under target shift. In *The 22nd AISTATS*, (pp. 849–858). PMLR.

Redko, I., Morvant, E., Habrard, A., Sebban, M., & Bennani, Y. (2019). *Advances in domain adaptation theory*. Elsevier.

Reich, S. (2013). A nonparametric ensemble transform method for bayesian inference. *SIAM Journal on Scientific Computing, 35*(4), A2013–A2024.

Saenko, K., Kulis, B., Fritz, M., & Darrell, T. (2010). Adapting visual category models to new domains. In *European conference on computer vision*, (pp. 213–226). Springer.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians* (pp. 55–58-63). Birkäuser.

Schmitzer, B., & Schnörr, C. (2013). A hierarchical approach to optimal transport. In *International conference on scale space and variational methods in computer vision*, (pp. 452–464). Springer.

Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10*(5), 1299–1319.

Shen, J., Qu, Y., Zhang, W., & Yu, Y. (2018). Wasserstein distance guided representation learning for domain adaptation. In *Thirty-second AAAI conference on artificial intelligence*.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8), 888–905.
Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference, 90*(2), 227–244.
Stella, X.Y., & Shi, J. (2003). Multiclass spectral clustering. In *Computer vision, IEEE international conference on*, (Vol 2, pp. 313–313). IEEE Computer Society.
Sturm, K. T. (2006). On the geometry of metric measure spaces. *Acta Mathematica, 196*(1), 65–131.
Sugiyama, M., Nakajima, S., Kashima, H., Von Buenau, P., & Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*.
Taherkhani, F., Dabouei, A., Soleymani, S., Dawson, J., & Nasrabadi, N.M. (2020). Transporting labels via hierarchical optimal transport for semi-supervised learning. In *ECCV*, (pp. 509–526).
Tsironis, S., Sozio, M., Vazirgiannis, M., & Poltechnique, L. (2013). Accurate spectral clustering for community detection in mapreduce. In *Advances in neural information processing systems (NIPS) workshops*, (p. 8). Citeseer.
Vapnik, V. (2013). *The nature of statistical learning theory*. Springer.
Venkateswara, H., Eusebio, J., Chakraborty, S., & Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 5018–5027).
Villani, C. (2009). *Optimal transport: old and new*. Springer.
Wilson, G., & Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST), 11*(5), 1–46.
Xu, R., Liu, P., Wang, L., Chen, C., & Wang, J. (2020). Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 4394–4403).
Yan, D., Huang, L., & Jordan, M.I. (2009). Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, (pp. 907–916).
Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F., & Solomon, J.M. (2019). Hierarchical optimal transport for document representation. In *Advances in neural information processing systems*.
Zha, H., He, X., Ding, C., Gu, M., & Simon, H.D. (2001). Spectral relaxation for k-means clustering. In *Advances in neural information processing systems*, (pp. 1057–1064).
Zhang, Y., Liu, T., Long, M., & Jordan, M. (2019). Bridging theory and algorithm for domain adaptation. In *International conference on machine learning*, (pp. 7404–7413). PMLR.