TIDOT: A Teacher Imitation Learning Approach for Domain Adaptation with Optimal Transport

Tuan Nguyen[1] , **Trung Le**[1] , **Nhan Dam**[1,2] , **Quan Hung Tran**[3] , **Truyen Nguyen**[4] and **Dinh Phung**[1,2]
[1]Department of Data Science and AI, Monash University, Australia
[2]VinAI Research, Vietnam
[3]Adobe Research, San Jose, CA, USA
[4]University of Akron, USA
{tuan.ng, trunglm, nhan.dam}@monash.edu, qtran@adobe.com,
tn8@uakron.edu, dinh.phung@monash.edu

Abstract

Using the principle of imitation learning and the theory of optimal transport we propose in this paper a novel model for unsupervised domain adaptation named Teacher Imitation Domain Adaptation with Optimal Transport (TIDOT). Our model includes two cooperative agents: a teacher and a student. The former agent is trained to be an expert on labeled data in the source domain, whilst the latter one aims to work with unlabeled data in the target domain. More specifically, optimal transport is applied to quantify the total of the distance between embedded distributions of the source and target data in the joint space, and the distance between predictive distributions of both agents, thus by minimizing this quantity TIDOT could mitigate not only the data shift but also the label shift. Comprehensive empirical studies show that TIDOT outperforms existing state-of-the-art performance on benchmark datasets.

1 Introduction

Domain adaptation or covariate shift problem has emerged from the observation of significant degradation in predictive performance when there exists a shift between a source domain (over which a classifier is trained) and a target domain (over which the classifier does prediction). Many state-of-the-art methods have been proposed for both shallow domain adaptation [Courty *et al.*, 2017b] and deep domain adaptation [Ganin and Lempitsky, 2015; Long *et al.*, 2015; French *et al.*, 2018; Shu *et al.*, 2018; Damodaran *et al.*, 2018].

Imitation learning follows the principle of *'learning from demonstration'*. In particular, there are two fundamental components: an expert teacher and a student. The former component knows how to do its job perfectly, whilst the latter one tries to imitate the teacher to solve its task. This learning principle has been applied successfully in reinforcement learning and sequence prediction [Abbeel and Ng, 2004; Ross *et al.*, 2011; Ho and Ermon, 2016].

In this work, using the principle of imitation learning and the theory of optimal transport we propose a novel model for unsupervised domain adaptation, named ***T**eacher **I**mitation **D**omain Adaptation with **O**ptimal **T**ransport* (**TIDOT**). The teacher in this scenario is apparently a classifier trained on the labeled source domain, but two questions naturally arise: i) which component is the student; and ii) what are the principle and mechanism to enable the student to mimic its teacher in this specific application? We address these two questions by developing a rigorous and intuitive theory based on the theory of optimal transport (OT) [Villani, 2008; Santambrogio, 2015; Peyré *et al.*, 2019]. From an abstract interpretation, our theory and mechanism postulate that to predict an unlabeled target sample the student needs to match this target sample with a corresponding labeled source sample so as to conveniently imitate the prediction of the teacher on this source sample. We summarize our contributions in this work as follows:

- We propose a rigorous OT-based theory to leverage imitation learning and domain adaptation. This paradigm is sufficiently general to potentially and promisingly apply to many learning problems including reinforcement learning. In this paper, we demonstrate its application in the context of unsupervised domain adaptation.
- We conduct experiments to compare our TIDOT to the baselines, especially OT-based deep DA, e.g., DeepJDOT [Damodaran *et al.*, 2018], SWD [Lee *et al.*, 2019], DASPOT [Xie *et al.*, 2019], ETD [Li *et al.*, 2020] and RWOT [Xu *et al.*, 2020]). The experimental results show that our proposed method outperforms existing methods on a variety of benchmark datasets including *Digits*, *traffic sign*, *natural scenes*, *Office-31*, *Office-Home*, and *ImageCLEF-DA*.
- We empirically suggest a potential OT-inspired regularization technique for future work. In particular, as an intriguing side effect of our proposed model, by setting the target training set as the source validation set, we en-

This work was supported by the US Air Force grant FA2386-19-1-4040.

force the teacher to not only predict well on the source training set, but also generalize to predict well on the unlabeled source validation set. We demonstrate that this strategy can yield a regularizer to mitigate the overfitting problem. Although this point is not the main claim of this work and we investigate it as an ablation study, its promising results reveal that this workaround is potentially a decent OT-inspired regularization technique.

2 Related Work

Deep domain adaptation (DA) has been intensively studied and shown appealing performance in various tasks and applications, notably [Ganin and Lempitsky, 2015; Long *et al.*, 2015; French *et al.*, 2018; Damodaran *et al.*, 2018; Nguyen *et al.*, 2019; 2020]. The core idea of deep DA is to bridge the gap between source and target distributions in a joint space by minimizing a divergence between distributions induced from the source and target domains. Popular choices of divergence include Jensen-Shannon divergence [Ganin and Lempitsky, 2015; Long *et al.*, 2015; French *et al.*, 2018]; maximum mean discrepancy distance [Long *et al.*, 2015]; and WS distance [Courty *et al.*, 2017b; Nguyen *et al.*, 2021; Le *et al.*, 2021].

Optimal transport theory has been applied in domain adaptation in [Courty *et al.*, 2017b; 2017a; Damodaran *et al.*, 2018; Redko *et al.*, 2019; Lee *et al.*, 2019; Xie *et al.*, 2019; Li *et al.*, 2020; Xu *et al.*, 2020]. Particularly, [Lee *et al.*, 2019] proposes using sliced Wasserstein distance for domain adaption, whereas [Xie *et al.*, 2019] introduces SPOT in which the optimal transport plan is approximated by a pushforward of a reference distribution, and cast the optimal transport problem into a minimax problem. Recently, ETD [Li *et al.*, 2020] measures the domain discrepancy by minimizing attention-aware transport distance while RWOT [Xu *et al.*, 2020] exploits spatial prototypical information and intra-domain structure to reduce the negative transfer brought by the target samples near decision boundaries. Moreover, [Courty *et al.*, 2017b] proposes a brilliant idea to connect the theory of optimal transport and domain adaptation [Courty *et al.*, 2017a], which later inspires an OT-based deep DA method (DeepJDOT) [Damodaran *et al.*, 2018] and learning from multiple data sources [Redko *et al.*, 2019]. Different from [Courty *et al.*, 2017b], our theory originates from OT-based imitation learning for which we develop a rigorous theory to explain the intuition of OT-based imitation DA and also theoretically analyze the general loss of OT-based imitation deep DA, wherein we employ deep neural networks for transfer learning. We note that this makes our theory significantly distinguish from [Courty *et al.*, 2017b] which only limits in the standard setting of transfer learning. In terms of modeling, we propose TIDOT which encourages OT-based imitation learning via a teacher and student in which the teacher guides and offers pseudo labels to the student, whereas the student tries to imitate the teacher. Furthermore, we invoke the clustering view of OT as an intuitive tool to explain why TIDOT can mitigate the label shift problem. Last but not least, our OT-based imitation learning viewpoint together with its developed theory is potential to apply to a broader context such as adversarial machine learning (AML), generative models, and imitation learning in reinforcement learning.

3 Related Background

3.1 OT with Entropic Regularized Duality

Consider two distributions $\mathbb{P}$ and $\mathbb{Q}$ which operate on the domain $\Omega \subseteq \mathbb{R}^d$, let $d(\mathbf{x}, \mathbf{y})$ be a non-negative and continuous cost function or metric. Wasserstein distance [Santambrogio, 2015; Villani, 2008] between $\mathbb{P}$ and $\mathbb{Q}$ w.r.t the metric d is defined as

$$\mathcal{W}_d(\mathbb{Q}, \mathbb{P}) := \min_{\gamma \in \Gamma(\mathbb{Q}, \mathbb{P})} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}\left[d(\mathbf{x}, \mathbf{y})\right], \tag{1}$$

where γ is a coupling that admits $\mathbb{Q}, \mathbb{P}$ as its marginals.

To enable the application of optimal transport in machine learning and deep learning, [Genevay *et al.*, 2016] developed an entropic regularized dual form. First, they proposed to add an entropic regularization term to primal form (1)

$$\mathcal{W}_d^\epsilon(\mathbb{Q}, \mathbb{P}) := \min_{\gamma \in \Gamma(\mathbb{Q}, \mathbb{P})} \left\{\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\gamma}\left[d(\mathbf{x}, \mathbf{y})\right] + \epsilon D_{KL}(\gamma \| \mathbb{Q} \otimes \mathbb{P})\right\}, \tag{2}$$

where ϵ is the regularization rate, $D_{KL}(\cdot\|\cdot)$ is the Kullback-Leibler (KL) divergence, and $\mathbb{Q} \otimes \mathbb{P}$ represents the specific coupling in which $\mathbb{Q}$ and $\mathbb{P}$ are independent.

Using Fenchel-Rockafellar theorem, they obtained the following *entropic regularized dual form* of (2)

$$\begin{aligned}\mathcal{W}_d^\epsilon(\mathbb{Q}, \mathbb{P}) &= \max_{\phi} \left\{\int \phi_\epsilon^c(\mathbf{x})\, \mathrm{d}\mathbb{Q}(\mathbf{x}) + \int \phi(\mathbf{y})\, \mathrm{d}\mathbb{P}(\mathbf{y})\right\} \\ &= \max_{\phi} \left\{\mathbb{E}_{\mathbb{Q}}\left[\phi_\epsilon^c(\mathbf{x})\right] + \mathbb{E}_{\mathbb{P}}\left[\phi(\mathbf{y})\right]\right\}, \end{aligned} \tag{3}$$

where $\phi_\epsilon^c(\mathbf{x}) := -\epsilon \log\left(\mathbb{E}_{\mathbb{P}}\left[\exp\left\{\frac{-d(\mathbf{x},\mathbf{y})+\phi(\mathbf{y})}{\epsilon}\right\}\right]\right)$.

3.2 Clustering View of Optimal Transport

This view of optimal transport has been utilized to study a rich class of hierarchical and multilevel clustering problems [Ho *et al.*, 2019; 2017]. We now present the clustering view of optimal transport which assists us to interpret our method developed in the sequel. Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete distributions defined as

$$\mathbb{P} := \tfrac{1}{m}\textstyle\sum_{i=1}^m \delta_{\mathbf{u}_i} \text{ and } \mathbb{Q} := \tfrac{1}{n}\textstyle\sum_{j=1}^n \delta_{\mathbf{v}_j},$$

where $\delta_{\mathbf{x}}$ indicates a Dirac measure centered at $\mathbf{x}$.

The clustering view reveals that if we learn the atoms of $\mathbb{Q}$ to minimize $\mathcal{W}_d(\mathbb{P}, \mathbb{Q})$ w.r.t the metric d, the optimal atoms of $\mathbb{Q}$ become the centroids of the clusters formed by the atoms of $\mathbb{P}$ or the atoms of $\mathbb{Q}$ are moving to find the groups of atoms of $\mathbb{P}$ with the aim to minimize the distortion w.r.t the metric d (see our supplementary material for more detail).

4 Our Proposed Method

4.1 Optimal Transport Based Imitation Learning

In what follows, we present OT-based imitation learning which lays foundation for our proposed TIDOT. Consider

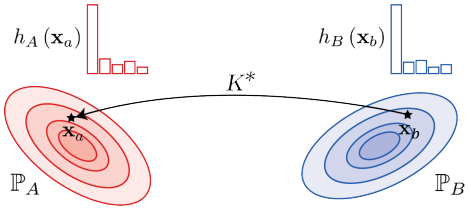

Figure 1: Imitation view explanation. $h_B(\mathbf{x}_b)$ for $\mathbf{x}_b \sim \mathbb{P}_B$ tries to imitate the prediction of $h_A(\mathbf{x}_a)$ for $\mathbf{x}_a = K^*(\mathbf{x}_b) \sim \mathbb{P}_A$.

two data domains $\mathcal{X}_A$ and $\mathcal{X}_B$ with two data distributions $\mathbb{P}_A$ and $\mathbb{P}_B$ respectively, we assume that $h_A : \mathcal{X}_A \to \mathcal{Y}_\triangle$ (where $\mathcal{Y}_\triangle := \left\{\boldsymbol{\pi} \in \mathbb{R}^M : \|\boldsymbol{\pi}\|_1 = 1 \text{ and } \boldsymbol{\pi} \geq \mathbf{0}\right\}$ and M is the number of classes), which is a well-qualified classifier that gives accurate prediction for data instances on $\mathcal{X}_A$ sampled from $\mathbb{P}_A$. We wish to learn a classifier h_B to predict accurately data instances sampled from $\mathbb{P}_B$ by imitating what is done by h_A on $(\mathcal{X}_A, \mathbb{P}_A)$.

To serve the development of OT-based imitation learning, given two pairs $\mathbf{z}_1 = \left(\mathbf{x}_1, y_1^\triangle\right) \in \mathcal{X}_S \times \mathcal{Y}_\triangle$ and $\mathbf{z}_2 = \left(\mathbf{x}_2, y_2^\triangle\right) \in \mathcal{X}_T \times \mathcal{Y}_\triangle$, we define the cost (distance) function between them as:

$$d(\mathbf{z}_1, \mathbf{z}_2) := \lambda d_x(\mathbf{x}_1, \mathbf{x}_2) + d_y\left(y_1^\triangle, y_2^\triangle\right), \tag{4}$$

where d_x is ground metric (cost) defined on $\mathcal{X}_S \times \mathcal{X}_T$ and d_y is a divergence defined on $\mathcal{Y}_\triangle$.

Based on the data distribution $\mathbb{P}_A$ and classifier h_A, we define a distribution $\mathbb{P}_{A,h_A}$ over $\mathcal{X}_A \times \mathcal{Y}_\triangle$ including sample pair $(\mathbf{x}, h_A(\mathbf{x}))$ by first sampling $\mathbf{x} \sim \mathbb{P}_A$ and then computing $h_A(\mathbf{x})$. Similarly, we can define another distribution $\mathbb{P}_{B,h_B}$ over $\mathcal{X}_B \times \mathcal{Y}_\triangle$ using the data distribution $\mathbb{P}_B$ and the classifier h_B. To allow h_B to imitate the behavior of h_A, we propose to inspect Wasserstein distance between $\mathbb{P}_{A,h_A}$ and $\mathbb{P}_{B,h_B}$ w.r.t the cost (metric) function d defined as above. The following proposition is crucial for us to speculate what it really means by OT-based imitation learning.

Proposition 1. *The WS distance of interest* $\mathcal{W}_d(\mathbb{P}_{A,h_A}, \mathbb{P}_{B,h_B})$ *can be expressed as:*

$$\min_{L:L_\#\mathbb{P}_A=\mathbb{P}_B} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_A}\left[\lambda d_x(\mathbf{x}, L(\mathbf{x})) + d_y(h_A(\mathbf{x}), h_B(L(\mathbf{x})))\right] =$$
$$\min_{K:K_\#\mathbb{P}_B=\mathbb{P}_A} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_B}\left[\lambda d_x(\mathbf{x}, K(\mathbf{x})) + d_y(h_B(\mathbf{x}), h_A(K(\mathbf{x})))\right].$$

According to Proposition 1, when computing $\mathcal{W}_d(\mathbb{P}_{A,h_A}, \mathbb{P}_{B,h_B})$, we need to find an optimal transport $K^* : K^*_\#\mathbb{P}_B = \mathbb{P}_A$ that moves $\mathbb{P}_B$ to $\mathbb{P}_A$ so as to minimize the difference in predictions of h_B and h_A. This further implies that given $\mathbf{x} \sim \mathbb{P}_B$, the prediction behavior of $h_B(\mathbf{x})$ imitates that of $h_A(K^*(\mathbf{x}))$ for $K^*(\mathbf{x}) \sim \mathbb{P}_A$. Figure 1 provides an intuitive explanation for our proposed imitation learning viewpoint based on optimal transport.

In addition, from Proposition 1, it is obvious that

$$\mathcal{W}_d(\mathbb{P}_{A,h_A}, \mathbb{P}_{B,h_B}) \geq \min_{K:K_\#\mathbb{P}_B=\mathbb{P}_A} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_B}\left[\lambda d_\mathcal{X}(\mathbf{x}, K(\mathbf{x}))\right]$$
$$= \lambda \mathcal{W}_{d_X}(\mathbb{P}_A, \mathbb{P}_B). \tag{5}$$

$$\mathcal{W}_d(\mathbb{P}_{A,h_A}, \mathbb{P}_{B,h_B}) \geq \min_{K:K_\#\mathbb{P}_B=\mathbb{P}_A} \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_B}\left[d_\mathcal{Y}(h_B(\mathbf{x}), h_A(K(\mathbf{x})))\right]. \tag{6}$$

In the context of unsupervised domain adaptation, $\mathbb{P}_B$ defined on a latent space via a feature extractor (generator) is hence an unfixed distribution, whereas h_B is a trainable classifier. Therefore, from Inequalities (5) and (6), when minimizing $\mathcal{W}_d(\mathbb{P}_{A,h_A}, \mathbb{P}_{B,h_B})$, we minimize $\mathcal{W}_{d_X}(\mathbb{P}_A, \mathbb{P}_B)$ to reduce the data shift between two data distributions and simultaneously find the transport map K to allow $h_B(\mathbf{x})$ (target example $\mathbf{x} \sim \mathbb{P}_B$) imitating $h_A(K(\mathbf{x}))$ ($K(\mathbf{x}) \sim \mathbb{P}_A$) for mitigating the label shift between two domains.

4.2 OT-based Imitation Learning on Domain Adaptation

In what follows, we present how to apply our OT-based imitation learning mechanism to unsupervised domain adaptation (UDA). The UDA setting include two datasets, a labeled dataset $\mathbb{D}_S = \left\{\left(\mathbf{x}_i^S, y_i^S\right)\right\}_{i=1}^{N_S}$ from a source domain with $\mathbf{x}_i^S \in \mathbb{R}^d$ and $y_i^S \in \{1, 2, ..., M\}$ and an unlabeled dataset $\mathbb{D}_T = \left\{\mathbf{x}_i^T\right\}_{i=1}^{N_T}$ from a target domain. We denote $\mathbb{P}_S$ and $\mathbb{P}_T$ as the empirical data distributions for the source and target domains, i.e., $\mathbb{P}_S = \frac{1}{N_S}\sum_{i=1}^{N_S}\delta_{\mathbf{x}_i^S}$ and $\mathbb{P}_T = \frac{1}{N_T}\sum_{i=1}^{N_T}\delta_{\mathbf{x}_i^T}$, where $\delta_\mathbf{x}$ indicates a Dirac measure centered at $\mathbf{x}$.

Following [Ganin and Lempitsky, 2015; Long *et al.*, 2013; Long *et al.*, 2015], we employ a generator (feature extractor) G to map both source and target examples into a latent space. On the latent space, we train a *teacher classifier* h_S using the labeled source dataset $\mathbb{D}_S$ and a *student classifier* h_T using our proposed OT-based imitation mechanism to reduce the data and label shifts. More specifically, we propose minimizing the objective function consisting of the empirical loss of the teacher h_S and an OT-based imitation learning term involving both the teacher h_S and the student h_T:

$$\min_{h_S, h_T, G}\left\{\mathcal{L}^S + \alpha\mathcal{R}^{WS}\right\}, \tag{7}$$

where $\alpha > 0$ is a trade-off parameter and we have defined

$$\mathcal{L}^S = \frac{1}{N_S}\sum_{i=1}^{N_S}\ell\left(h_S\left(G\left(\mathbf{x}_i^S\right)\right), y_i^S\right),$$

for which ℓ is a loss function (e.g., the cross-entropy loss) and

$$\mathcal{R}^{WS} = \mathcal{W}_d(\mathbb{P}_{T,h_T}, \mathbb{P}_{S,h_S}),$$

for which $\mathbb{P}_{S,h_S}$ is the joint distribution constituted by the pairs $(G(\mathbf{x}), h_S(G(\mathbf{x})))$ where $\mathbf{x} \sim \mathbb{P}^S$ and $\mathbb{P}_{T,h_T}$ is the joint distribution constituted by the pairs $(G(\mathbf{x}), h_T(G(\mathbf{x})))$ where $\mathbf{x} \sim \mathbb{P}_T$. Note that the ground metric (cost) d now involves the latent space and is defined as:

$$d(\mathbf{z}_1, \mathbf{z}_2) = \lambda d_x(G(\mathbf{x}_1), G(\mathbf{x}_2)) + d_y(h_S(G(\mathbf{x}_1)), h_T(G(\mathbf{x}_2))), \tag{8}$$

where $\mathbf{z}_1 = (G(\mathbf{x}_1), h_S(G(\mathbf{x}_1)))$ and $\mathbf{z}_2 = (G(\mathbf{x}_2), h_T(G(\mathbf{x}_2)))$ and $d_\mathcal{X}(\cdot,\cdot)$ is a distance between two data examples on the latent space.

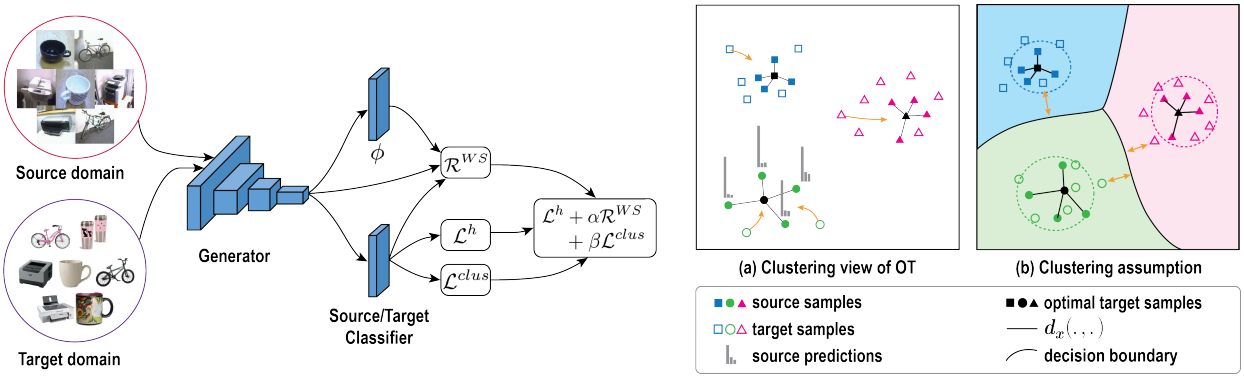

Figure 2: The overall structure of the TIDOT method for unsupervised domain adaptation. Our training model consists of three main components, namely a weight-sharing generator, classifiers of source and target, and Kantorovich potential network (ϕ). Via the generator, source and target examples are mapped into a latent space where the classifiers and ϕ are acted on it. The output of the generator, classifiers and ϕ are used for computing losses $\mathcal{L}^h$, $\mathcal{R}^{WS}$ and $\mathcal{L}^{clus}$. Minimizing cross-entropy loss $\mathcal{L}^h$ ensures the model predicts well on source examples, whereas $\mathcal{R}^{WS}$ and $\mathcal{L}^{clus}$ significantly contribute to domain adaptation: (a) On the latent space, target samples try to find an appropriate cluster of source samples with guarantees of OT-based clustering view; (b) Our proposal leverages the cluster assumption to encourage the classifier to be more confident on target samples lying on the decision boundary. *Best viewed in color*.

Generally, we train a teacher classifier h_S to predict well on the source domain, while using the OT-based imitation mechanism to train a student classifier h_T to move target representations to source representations on the latent space (i.e., reducing data shift), whereas encouraging h_T to mimic the predictions of h_S (i.e., reducing label shift). In our experiments, the student classifier h_T consistently outperforms the teacher classifiers. The reason is possibly that though target representations tend to move to source representations, there always exists a gap, hence the teacher classifier h_S trained on source examples is hard to predict perfectly target examples, whereas the student classifier performs better because it is trained on target examples via imitating the predictions of h_S on relevant source examples.

Clustering View Explanation. We now explain why minimizing $\mathcal{R}^{WS} = \mathcal{W}_d\left(\mathbb{P}_{T,h_T}, \mathbb{P}_{S,h_S}\right)$ can help to mitigate the data shift and label shift, two thorny issues existing in UDA. This is intuitively explainable from the *clustering view of the optimal transport* (see Section 3.2). More specifically, let us denote $\mathbf{z}_i^S = \left(G\left(\mathbf{x}_i^S\right), h_S\left(G\left(\mathbf{x}_i^S\right)\right)\right), \forall i = 1, ..., N_S$ and $\mathbf{z}_i^T = \left(G\left(\mathbf{x}_i^T\right), h_T\left(G\left(\mathbf{x}_i^T\right)\right)\right), \forall i = 1, ..., N_T$. It appears

$$\mathbb{P}_{S,h_S} = \frac{1}{N_S}\sum_{i=1}^{N_S}\delta_{\mathbf{z}_i^S} \text{ and } \mathbb{P}_{T,h_T} = \frac{1}{N_T}\sum_{i=1}^{N_T}\delta_{\mathbf{z}_i^T}.$$

Referred to the clustering view of OT, when minimizing $\mathcal{R}^{WS} = \mathcal{W}_d\left(\mathbb{P}_{T,h_T}, \mathbb{P}_{S,h_S}\right)$, we encourage each $\mathbf{z}_i^T$ in the target domain to find an appropriate group (cluster) of $\mathbf{z}_j^S$ with $j \in J$ for some set of indices $J \subset \{1, ..., N_S\}$ so that the total distortion $\sum_{j\in J} d\left(\mathbf{z}_i^T, \mathbf{z}_j^S\right)$ defined as

$$\sum_{j\in J}\Big(\lambda d_x\left(G\left(\mathbf{x}_i^T\right), G\left(\mathbf{x}_j^S\right)\right) + d_y\left(h_T\left(G\left(\mathbf{x}_i^T\right)\right), h_S\left(G\left(\mathbf{x}_j^S\right)\right)\right)\Big)$$

is minimized. This further implies that *(i)* $G\left(\mathbf{x}_i^T\right)$ *moves toward the group or cluster of* $G\left(\mathbf{x}_j^S\right), j \in J$ and *(ii) the predictions of* h_S *for the source data examples in that group or cluster (i.e.,* $h_S\left(G\left(\mathbf{x}_j^S\right)\right), j \in J$*) need to be a consensus*. The first conclusion (i) is straight-forward from minimizing $d_x\left(G\left(\mathbf{x}_i^T\right), G\left(\mathbf{x}_j^S\right)\right)$, whilst the second conclusion (ii) comes from the fact that the prediction $h_T\left(G\left(\mathbf{x}_i^T\right)\right)$ mimics those of $h_S\left(G\left(\mathbf{x}_j^S\right)\right)$ for all $j \in J$, hence $h_S\left(G\left(\mathbf{x}_j^S\right)\right), \forall j \in J$ should reach a consensus on their predictions. Eventually, each $G\left(\mathbf{x}_i^T\right)$ is encouraged to move to a group or cluster of $G\left(\mathbf{x}_j^S\right), j \in J$ in the source domain which shares the same prediction label to imitate their common prediction. This would help to mitigate the label shift issue (see Figure 2).

4.3 Entropic Regularized Solution

To tackle the OT-based regularization term $\mathcal{R}^{WS}$, we use *entropic regularized duality form* (see Eq. (3)) of optimal transport. Specifically, we approximate $\mathcal{R}^{WS} = \mathcal{W}_d^{\epsilon}\left(\mathbb{P}_{T,h_T}, \mathbb{P}_{S,h_S}\right)$ which has the following form:

$$\mathcal{R}^{WS} = \max_{\phi}\Bigg\{\frac{1}{N_T}\sum_{i=1}^{N_T}\Bigg[-\epsilon\log\Bigg(\frac{1}{N_S}\sum_{j=1}^{N_S}\exp\Bigg\{\frac{1}{\epsilon}\Big[\phi\left(G\left(\mathbf{x}_j^S\right)\right) - d\left(\mathbf{z}_i^T, \mathbf{z}_j^S\right)\Big]\Bigg\}\Bigg)\Bigg] + \frac{1}{N_S}\sum_{j=1}^{N_S}\phi\left(G\left(\mathbf{x}_j^S\right)\right)\Bigg\}, \quad (9)$$

where $d\left(\mathbf{z}_i^T, \mathbf{z}_j^S\right) = \lambda d_x\left(G\left(\mathbf{x}_i^T\right), G\left(\mathbf{x}_j^S\right)\right) + d_y\left(h_T\left(G\left(\mathbf{x}_i^T\right)\right), h_S\left(G\left(\mathbf{x}_j^S\right)\right)\right)$ is the transportation cost, ϕ is a neural net named Kantorovich potential network (see Eq. (3)).

4.4 Ensuring Clustering Assumption for TIDOT

Clustering assumption [Chapelle and Zien, 2005] is a technique that encourages the classifier to preserve its predictions for data examples in a cluster. Basically, the clustering assumption enforces the decision boundary of a given classifier

to lie in the gap among the data clusters and never crosses over any clusters. We observe that coupling the clustering assumption for the classifiers h_S, h_T with TIDOT helps to boost its performance.

The reason for this complementary collaboration is that although minimizing the OT-based regularization term $\mathcal{R}^{WS}$ helps to move the target example $G\left(\mathbf{x}_i^T\right)$ to a group or cluster of the source examples $G\left(\mathbf{x}_j^S\right)$ with the same label, the prediction of $h_T\left(G\left(\mathbf{x}_i^T\right)\right)$ as in Eq. (9) is encouraged to mimic the predictions of $h_S\left(G\left(\mathbf{x}_j^S\right)\right)$ including diverge data examples of different classes. Therefore, the prediction of $h_T\left(G\left(\mathbf{x}_i^T\right)\right)$ for those lying on the cluster boundary tends to be possibly less confident and misleading.

With the assistance of the clustering assumption, the classifier h_T is strengthened to predict well the target examples lying on the clustering boundary. Specifically, this encourages the classifier h_T to predict those target examples using the same label as others in the cluster, hence correcting the predictions for those examples. To enforce the clustering assumption, we employ Virtual Adversarial Training (VAT) [Miyato *et al.*, 2019] in conjunction with minimizing the entropy of prediction [Grandvalet and Bengio, 2005] as in [Shu *et al.*, 2018; Kumar *et al.*, 2018]

$$\mathcal{L}^{clus} = \mathcal{L}^{ent} + \mathcal{L}^{vat},$$

where with $\mathbb{H}$ to be the entropy, we have defined

$$\begin{aligned}\mathcal{L}^{ent} &= \mathbb{E}_{\mathbb{P}_\mathbb{T}}\left[\mathbb{H}\left(h_T\left(G\left(\mathbf{x}\right)\right)\right)\right], \\ \mathcal{L}^{vat} &= \mathbb{E}_{\mathbb{P}_\mathbb{S}}\left[\max_{\mathbf{x}':\|\mathbf{x}'-\mathbf{x}\|<\theta} D_{KL}\left(h_S\left(G\left(\mathbf{x}\right)\right), h_S\left(G\left(\mathbf{x}'\right)\right)\right)\right] \\ &\quad + \mathbb{E}_{\mathbb{P}_T}\left[\max_{\mathbf{x}':\|\mathbf{x}'-\mathbf{x}\|<\theta} D_{KL}\left(h_T\left(G\left(\mathbf{x}\right)\right), h_T\left(G\left(\mathbf{x}'\right)\right)\right)\right]\end{aligned}$$

with which D_{KL} represents a Kullback-Leibler divergence and θ is a very small positive number.

4.5 Teacher Imitation Domain Adaptation Based on Optimal Transport

The final optimization problem of our TIDOT is as follows:

$$\min_{h_S, h_T, G}\left\{\mathcal{L}^h + \alpha\mathcal{R}^{WS} + \beta\mathcal{L}^{clus}\right\}, \tag{10}$$

where $\beta > 0$ is a trade-off parameter. Under the clustering view of OT, it is worth noting that when minimizing $\mathcal{R}^{WS} = \mathcal{W}_d^\epsilon\left(\mathbb{P}_{T,h_T}, \mathbb{P}_{S,h_S}\right)$ in Eq. (10), we aim to push the representations of source and target data to be intermingled in the joint space and encourage each $G\left(\mathbf{x}_i^T\right)$ to find its corresponding $G\left(\mathbf{x}_j^S\right)$ to mimic the prediction of h_S. Finally, the pseudocode for the training process of TIDOT is presented in Algorithm 1 which is placed in our supplementary material due to the space limitation.

5 Experiments

5.1 Model Evaluation

In this section, we conduct experiments on four main datasets to evaluate our TIDOT with state-of-the-art domain adaptation methods: (1) ResNet-50 [He *et al.*, 2016]; (2) DANN [Ganin and Lempitsky, 2015]; (3) Π-model [French *et al.*, 2018]; (4) iCAN [Zhang *et al.*, 2018]; (5) CDAN [Long *et al.*, 2018]; (6) SWD [Lee *et al.*, 2019]; (7) DeepJDOT [Damodaran *et al.*, 2018]; (8) DASPOT [Xie *et al.*, 2019]; (9) ETD [Li *et al.*, 2020]; (10) RWOT [Xu *et al.*, 2020].

Method	MN US	US MN	MN MM	SV MN	MN SV	SS GT	CI ST	ST CI
DANN	-	-	81.5	71.1	35.7	88.7	-	-
Π-model	-	-	-	92.0	71.4	98.4	76.3	64.2
CDAN	95.6	98.0	-	89.2	-	-	-	-
SWD	98.1	97.1	-	98.9	-	98.6	-	-
DeepJDOT	95.7	96.4	92.4	96.7	-	-	-	-
DASPOT	97.5	96.5	94.9	96.2	-	-	-	-
RWOT	**98.5**	97.5	-	98.8	-	-	-	-
TIDOT teacher	98.1	98.6	97.7	98.9	82.2	98.8	76.5	72.3
TIDOT student	98.3	**99.0**	**98.5**	**99.0**	**86.8**	**99.1**	**77.4**	**75.0**

Table 1: Classification accuracy (%) on Digits, traffic sign and natural image datasets.

Digits, Traffic Sign, and Natural Scenes Datasets include MNIST (**MN**), USPS (**US**), MNIST-M (**MM**), Synthetic Digits (**SN**), Street View House Numbers (**SV**), Synthetic Traffic Signs (**SS**), German Traffic Signs Recognition Benchmark (**GT**), CIFAR-10 (**CI**), and STL-10 (**ST**).

Office-31 contains 3 domains Amazon (**A**), Webcam (**W**), and DSLR (**D**). There are 31 common classes for all domains and the total number of images is 4,110.

Office-Home consists of roughly 15,500 images in a total of 65 object classes and belonging to 4 different domains: Artistic (**Ar**), Clip Art (**Cl**), Product (**Pr**) and Real-world (**Rw**).

ImageCLEF-DA contains three domains: Caltech-256 (**C**), ImageNet ILSVRC 2012 (**I**) and Pascal VOC 2012 (**P**). There are total 600 images in each domain and 12 common classes.

Method	A→W	A→D	D→W	W→D	D→A	W→A	Avg
ResNet-50	70.0	65.5	96.1	99.3	62.8	60.5	75.7
DANN	81.5	74.3	97.1	99.6	65.5	63.2	80.2
iCAN	92.5	90.1	98.8	**100.0**	72.1	69.9	87.2
CDAN	94.1	92.9	98.6	**100.0**	71.0	69.3	87.7
DeepJDOT	88.9	88.2	98.5	99.6	72.1	70.1	86.2
ETD	92.1	88.0	**100.0**	**100.0**	71.0	67.8	86.2
RWOT	95.1	94.5	99.5	**100.0**	77.5	77.9	90.8
TIDOT teacher	94.3	95.1	97.6	99.8	86.6	84.5	93.0
TIDOT student	**96.2**	**96.4**	98.1	**100.0**	**88.1**	**85.9**	**94.1**

Table 2: Classification accuracy (%) on Office-31 dataset using either ResNet-50 features or ResNet-50 based deep models.

5.2 Implementation Detail

Architecture. We employ small, medium and large network architectures whose detail in the supplementary material. To compare with baselines on *Office-Home* and *Office-31*, all transfer tasks use the pre-trained ResNet-50 [He *et al.*, 2016] features which have 2,048 dimensions.

Hyperparameters. The specifications of hyperparameters are described in the supplementary material.

5.3 Result and Discussion

We first evaluate TIDOT on *Digits*, *traffic sign*, and *natural scene* datasets and report the results in Table 1. The experimental results show that whilst TIDOT teacher outperforms

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
RWOT	55.2	72.5	78.0	63.5	72.5	75.1	60.2	48.5	78.9	69.8	54.8	82.5	67.6
TIDOT teacher	55.6	74.3	81.2	65.9	77.4	78.3	60.2	52.8	79.9	70.9	56.8	82.0	69.6
TIDOT student	**55.9**	**77.7**	82.5	**67.2**	**78.2**	**79.7**	**61.0**	**54.8**	**81.4**	**71.0**	**58.0**	**83.4**	**70.9**

Table 3: Classification accuracy (%) on Office-Home dataset using ResNet-50 features.

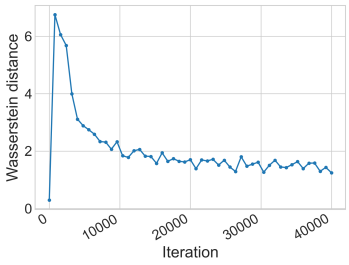

Figure 3: The visualization of $\mathcal{W}_d^\epsilon(\mathbb{P}_{T,h_T}, \mathbb{P}_{S,h_S})$ on **MN→SV** during the training process.

almost all state-of-the-art baselines, TIDOT student further enhances the accuracy of TIDOT teacher with various degrees of improvement (i.e., from as marginal as 0.1% to as significant as 4.6%). It is noticeable that although the transfer task **MN→SV** is extremely challenging in which the source dataset includes grayscale handwritten digits whereas the target dataset is created by real-world digits, our TIDOT is still capable of mitigating the shift of data and label between domains and outperforms the second-best method by a sizeable margin (15.4%).

We further testify TIDOT's performance on *Office-31* and report the classification results in Table 2. In general, our model achieves 94.1% on average and significantly outperforms on challenging adaptation where the source and target images are dissimilar in the background, i.e., **D→A**, **W→A**.

The results on *Office-Home* are reported in Table 3. TIDOT student exceeds almost comparison methods and achieves state-of-the-art performance, experiencing a go up by 3.3% on average compared with RWOT. More specifically, our model sees a remarkable improvement on challenging adaptation tasks, namely **Ar→Pr**, **Cl→Pr**.

The full comparison table for *Office-Home* and results on *ImageCLEF-DA* are shown in the supplementary material due to the limit of space.

5.4 Ablation Study

An Intuitive OT-based Loss. We further plot the values of $\mathcal{R}^{WS} = \mathcal{W}_d^\epsilon(\mathbb{P}_{T,h_T}, \mathbb{P}_{S,h_S})$ when simultaneously training teacher and student on the pair **MN→SV**. As shown in Figure 3, the cost of the optimal transport plan for moving from $\mathbb{P}_{T,h_S}$ to $\mathbb{P}_{S,h_S}$ is minimized, which means the data shift and label shift tends to be mitigated during the training process. Another remarkable advantage of OT-based methods as our TIDOT is that due to the effect of the envelope theorem, $\mathcal{R}^{WS}$ smoothly decreases while losses developed based on generative adversarial network (GAN) [Goodfellow *et al.*, 2014] always sees largely fluctuates.

Effect of Clustering Assumption. We investigate the effectiveness of the VAT loss ($\mathcal{L}^{vat}$) w.r.t. source and target distribution, and conditional entropy loss w.r.t. target distribution ($\mathcal{L}^{ent}$) on *ImageCLEF-DA*. The results in Table 4 show that by adding VAT loss (fourth row), the model sees a rise averagely by 1.4% compared to the basic setting (second row). Moreover, the figures are better when $\mathcal{L}^{vat}$ is combined with $\mathcal{L}^{ent}$ (fifth row). Via this ablation study, we find that VAT in conjunction with minimizing entropy supports our TIDOT to predict well on target samples lying on the decision boundary, and hence boots model performance further.

$\mathcal{L}^{vat}$	$\mathcal{L}^{ent}$	I→P	P→I	I→C	C→I	C→P	P→C	Avg
		79.5	90.5	96.7	93.5	78.7	95.8	89.1
	✓	80.2	91.0	96.8	93.2	79.5	96.0	89.5
✓		80.7	**93.8**	96.8	94.0	81.2	96.5	90.5
✓	✓	**81.7**	**93.8**	**97.5**	**94.5**	**81.2**	**96.6**	**90.8**

Table 4: Accuracy (%) of ablation study on ImageCLEF-DA.

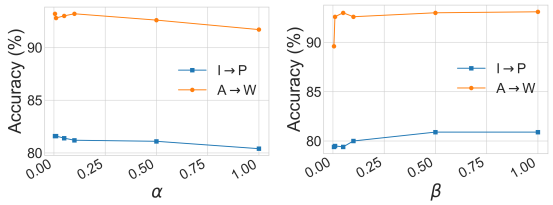

Figure 4: Analysis of model parameter w.r.t. α and β on **A→W** (orange line) and **I→P** (blue line).

Parameter Sensitivity. We further evaluate the effects of the trade-off parameters α, β in Figure 4. We search α, β in the grid of $\{0.005, 0.01, 0.05, 0.1, 0.5, 1.0\}$ and report the test accuracy on transfer tasks **A→W** and **I→P**. The results show that the model yields high performances when α from 0.005 to 0.1 and β from 0.05 to 1.0. However, with the other values of α and β, our model still achieves significant performances, which demonstrates the robustness and flexibility of TIDOT.

6 Conclusion

In this paper, we leverage the perspective of imitation learning and the theory of optimal transport to propose Teacher Imitation Domain Adaptation with Optimal Transport (TIDOT). Via two fundamental components of TIDOT, a teacher and a student, we apply our proposed method to unsupervised domain adaptation and conduct comprehensive experiments to compare TIDOT against the baselines. The experimental results show that our TIDOT outperforms the existing state-of-the-art OT-based method. Additionally, as a side effect of our developed theory, we interestingly discover a novel regularization technique for deep networks based on optimal transport, which is potential for future work.

References

[Abbeel and Ng, 2004] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.

[Chapelle and Zien, 2005] O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *AISTATS*, 2005.

[Courty *et al.*, 2017a] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *NIPS*, 2017.

[Courty *et al.*, 2017b] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE TPAMI*, 2017.

[Damodaran *et al.*, 2018] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *ECCV*, 2018.

[French *et al.*, 2018] G. French, M. Mackiewicz, and M. H. Fisher. Self-ensembling for visual domain adaptation. In *ICLR*, 2018.

[Ganin and Lempitsky, 2015] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

[Genevay *et al.*, 2016] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *NIPS*. 2016.

[Goodfellow *et al.*, 2014] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[Grandvalet and Bengio, 2005] Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *NIPS*. 2005.

[He *et al.*, 2016] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Ho and Ermon, 2016] J. Ho and S. Ermon. Generative adversarial imitation learning. In *NIPS*, 2016.

[Ho *et al.*, 2017] N. Ho, X. L. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung. Multilevel clustering via wasserstein means. In *ICML*, 2017.

[Ho *et al.*, 2019] N. Ho, V. Huynh, D. Phung, and M. I. Jordan. Probabilistic multilevel clustering via composite transportation distance. In *AISTATS*, 2019.

[Kumar *et al.*, 2018] A. Kumar, P. Sattigeri, K. Wadhawan, L. Karlinsky, R. Feris, B. Freeman, and G. Wornell. Co-regularized alignment for unsupervised domain adaptation. In *NeurIPS*. 2018.

[Le *et al.*, 2021] T. Le, T. Nguyen, N. Ho, H. Bui, and D. Phung. Lamda: Label matching deep domain adaptation. *ICML*, 2021.

[Lee *et al.*, 2019] C. Lee, T. Batra, M. H. Baig, and D. Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*, 2019.

[Li *et al.*, 2020] M. Li, Y. Zhai, Y. Luo, P. Ge, and C. Ren. Enhanced transport distance for unsupervised domain adaptation. In *CVPR*, 2020.

[Long *et al.*, 2013] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer feature learning with joint distribution adaptation. In *ICCV*, 2013.

[Long *et al.*, 2015] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

[Long *et al.*, 2018] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. In *NeurIPS*. 2018.

[Miyato *et al.*, 2019] T. Miyato, S. Maeda, M. Koyama, and S. Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE TPAMI*, 2019.

[Nguyen *et al.*, 2019] V. Nguyen, T. Le, T. Le, K. Nguyen, O. De Vel, P. Montague, L. Qu, and D. Phung. Deep domain adaptation for vulnerable code function identification. In *IJCNN*, 2019.

[Nguyen *et al.*, 2020] V. Nguyen, T. Le, O. De Vel, P. Montague, J. Grundy, and D. Phung. Dual-component deep domain adaptation: A new approach for cross project software vulnerability detection. In *PAKDD*, 2020.

[Nguyen *et al.*, 2021] T. Nguyen, T. Le, H. Zhao, H. Q. Tran, T. Nguyen, and D. Phung. Most: Multi-source domain adaptation via optimal transport for student-teacher learning. *UAI*, 2021.

[Peyré *et al.*, 2019] G. Peyré, M. Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 2019.

[Redko *et al.*, 2019] I. Redko, N. Courty, R. Flamary, and D. Tuia. Optimal transport for multi-source domain adaptation under target shift. In *AISTATS*, 2019.

[Ross *et al.*, 2011] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

[Santambrogio, 2015] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser*, 2015.

[Shu *et al.*, 2018] R. Shu, H. H. Bui, H. Narui, and S. Ermon. A DIRT-t approach to unsupervised domain adaptation. In *ICLR*, 2018.

[Villani, 2008] C. Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.

[Xie *et al.*, 2019] Y. Xie, M. Chen, H. Jiang, T. Zhao, and H. Zha. On scalable and efficient computation of large scale optimal transport. In *ICML*, 2019.

[Xu *et al.*, 2020] R. Xu, P. Liu, L. Wang, C. Chen, and J. Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *CVPR*, 2020.

[Zhang *et al.*, 2018] W. Zhang, W. Ouyang, W. Li, and D. Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *CVPR*, 2018.