Joint Partial Optimal Transport for Open Set Domain Adaptation

Renjun Xu[1*], **Pelen Liu**[1], **Yin Zhang**[1*], **Fang Cai**[2], **Jindong Wang**[3], **Shuoying Liang**[1], **Heting Ying**[4], **Jianwei Yin**[1]
[1]Zhejiang University
[2]Stanford University
[3]Microsoft Research
[4]China Jiliang University

Abstract

Domain adaptation (DA) has achieved a resounding success to learn a good classifier by leveraging labeled data from a source domain to adapt to an unlabeled target domain. However, in a general setting when the target domain contains classes that are never observed in the source domain, namely in Open Set Domain Adaptation (OSDA), existing DA methods failed to work because of the interference of the extra unknown classes. This is a much more challenging problem, since it can easily result in negative transfer due to the mismatch between the unknown and known classes. Existing researches are susceptible to misclassification when target domain unknown samples in the feature space distributed near the decision boundary learned from the labeled source domain. To overcome this, we propose Joint Partial Optimal Transport (JPOT), fully utilizing information of not only the labeled source domain but also the discriminative representation of unknown class in the target domain. The proposed joint discriminative prototypical compactness loss can not only achieve intra-class compactness and inter-class separability, but also estimate the mean and variance of the unknown class through backpropagation, which remains intractable for previous methods due to the blindness about the structure of the unknown classes. To our best knowledge, this is the first optimal transport model for OSDA. Extensive experiments demonstrate that our proposed model can significantly boost the performance of open set domain adaptation on standard DA datasets.

1 Introduction

Deep learning recently has improved the progress of diverse computer vision tasks such as image classification, object detection and semantic segmentation with the help of large-scale labeled datasets which is time and labor consuming to collect. However, it is often highly difficult to achieve satisfying performance when transferring knowledge from label-rich domain ($i.e.$, source domain) to label-scarce domain ($i.e.$, target domain), especially when there are unknown samples in the target domain. The characteristics of large scale unlabeled data in the target domain can be different from the labeled source domain, resulted in misclassification.

An underlying assumption in domain adaptation is that samples in the target domain necessarily belong to the classes observed in the source domain, from which we train the classifier. Since the target classes are often larger than training classes in real-world settings, considering the open set domain problem could lead to more practical results and applications. Under this regime, the target domain consists of subset of classes of the source domain and along with extra unknown classes. The main task is seeking to classify data of known classes correctly, and categorize the extra classes as "unknown". Lack of openness adaptation is a major critique against OSPB [Saito *et al.*, 2018].

We solve the problems above by a strategy based on the optimal transport framework. Optimal transport has been applied to closed set domain adaptation recently [S. and S., 2016; Luo *et al.*, 2017; Damodaran *et al.*, 2018]. Damodaran *et al.* recognized the strength of Convolutional Neural Network (CNN) to obtain a scalable solution by learning jointly the feature embedding between the two domains and the classifier in a single CNN framework. All previous OT based methods align the whole source and target domains and perform badly for OSDA since data of unknown classes in the target domain can make performance of domain adaptation model even inferior to a model without adaptation. Such phenomenon is known as *negative transfer* [Pan and Yang, 2010]. Unknown samples in the target domain need to be excluded in the transfer to avoid negative transfer. To this end, we propose a more flexible partial optimal transport framework that allows for transportation in only well-matched pairs of samples. We formulate the problem as transporting a fraction of the "well-matched" mass of source domain onto the target domain minimizing an overall displacement cost. We eliminate the far-fetched pairings that cause negative transfer in the global optimal coupling to obtain a partial optimal coupling. Besides the regular displacement loss, the overall displacement loss includes the loss for displacing to the unknown class applied to the fraction of mass not transferred. As a bonus, our method naturally adapts to different levels of openness of the target domain, because fraction size selected

*Corresponding authors: {rux, zhangyin98}@zju.edu.cn

from global coupling is data-adaptive.

Furthermore, we want to choose the feature representation that leads to better generalization and robust performance. For this goal, we include dispersion penalty for each class so that samples are concentrated around the means of their classes and therefore lie far from the decision boundary; we add to the objective function the joint discriminative prototypical compactness loss, which provides a way to estimate the mean and manipulate (shrink) the covariance of each class (including the unknown class). We firstly estimate the mean and variance for the unknown class from the residuals of our partial optimal transport which was a challenge for previous researchers as we are blind about the structure of unknown classes. Experimental results show that JPOT works stably in various datasets and outperforms existing methods.

2 Related Work

Closed Set Domain Adaptation (CSDA). Currently closed set domain adaptation methods [Tzeng *et al.*, 2014; Long *et al.*, 2015; Long *et al.*, 2016] minimized the feature distribution discrepancy to alleviate performance degradation in both source and target domains. Zellinger et al. [Zellinger *et al.*, 2017] achieved domain adaption by Central Moment Discrepancy (CMD). Contrastive Adaptation Network (CAN) [Kang *et al.*, 2019] optimizes a new metric that explicitly models the intra-class domain discrepancy and the inter-class domain discrepancy. Transferrable Prototypical Networks (TPN) [Pan *et al.*, 2019] exploits the prototypical distance [Snell *et al.*, 2017] to avoid misclassification.

Open Set Domain Adaptation. OSDA differs from CSDA by extra unknown classes in the target domain but not in the source domain. Recently, researchers are concentrating on open set recognition, intending to recognize "unknown" samples in the target domain during testing [Ganin and Lempitsky, 2015; Saito *et al.*, 2018]. Separate to Adapt (STA) network was proposed by [Liu *et al.*, 2019] to adapt different openness between source and target classes with coarse-to-fine weighting mechanism. Universal Adaptation Network [You *et al.*, 2019] was proposed to quantify sample-level transferability and recognize the "unknown" samples. However, all of them failed to solve the problem that target domain samples were distributed near the decision boundary learned from the source domain, leading to degraded performance.

Optimal Transport on Domain Adaptation. Optimal transport has been applied in domain adaptation to align the representations between the source domain and target domain [Courty *et al.*, 2016; Yan *et al.*, 2018]. As mentioned above, the coupling γ [Courty *et al.*, 2017] was applied to transport the source samples to the target domain by an estimated mapping. Deep Joint Optimal Transport (DeepJDOT) [Damodaran *et al.*, 2018] applied the coupling matrix γ to transport the source samples to the target domain in discriminative feature space, achieving high accuracy on many transfer tasks. However, the discriminative representation of "unknown" samples in the target domain has not been considered in previous approaches that causes mismatching.

3 Methodology

Following the settings of OSDA [Liu *et al.*, 2019], we define a *source* domain $\mathcal{D}^s = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ with n_s labeled samples and a *target* domain $\mathcal{D}^t = \{(\boldsymbol{x}_i^t)\}_{i=1}^{n_t}$ with n_t unlabeled samples. The label space in the source domain is defined as $\mathcal{C}^s$, which is a subspace of target-domain label space $\mathcal{C}^t$. The target domain further contains additional classes $\mathcal{C}^t_{unk}$, which are all referred as "unknown" since these classes are absent in the source domain, i.e. $\mathcal{C}^t = \mathcal{C}^s \cup \mathcal{C}^t_{unk}$. Data from source and target domains can have different probability distributions p and q respectively ($p \neq q$). In OSDA, we further observe that $p \neq q_{\mathcal{C}^s}$, where $q_{\mathcal{C}^s}$ represents the probability distribution of the target data belonging to shared classes $\mathcal{C}^s$.

Following [Saito *et al.*, 2018; Chen *et al.*, 2019], we adopt the two-stream CNN architecture with shared weights. We propose joint partial optimal transport (**JPOT**) for open set domain adaptation, which is an end-to-end method that learns a feature generator $G_f(\cdot)$ and a classifier $G_y(\cdot)$ to separate known and unknown classes in the target domain, as illustrated in Figure 2.

3.1 Partial Optimal Transport

Optimal transport for domain adaptation performs the alignment of the sample representations between the source domain and target domain [Courty *et al.*, 2016]. However, existing optimal transport strategies fail to address the problem caused by the pair-wise transport error in pairing unknown target samples to known source samples, resulting in negative transfer of unknown classes in the target domain. Despite similarities with the formulation [Damodaran *et al.*, 2018], we propose partial optimal transport strategy exploiting global coupling matrix γ given by joint probability distribution, which aims to identify the most correlated features of shared classes between the source and target domains. Moreover, we utilize global optimal matrix to capture distinct features of unknown class samples in the target domain and to realize inter-class separability of unknown vs known classes. The main procedure of JPOT is shown in the Figure 1.

The optimization of partial optimal transport is based on partial Kantorovich problem [Angenent *et al.*, 2003] seeking for a general coupling $\gamma \in \mathcal{X}(\mathcal{D}^s, \mathcal{D}^t)$ between $\mathcal{D}^s$ and $\mathcal{D}^t$:

$$\gamma^* = \underset{\gamma \in \mathcal{X}(\mathcal{D}^s, \mathcal{D}^t)}{\arg\min} \int_{\mathcal{D}^s \times \mathcal{D}^t} \mathcal{M}(\boldsymbol{x}^s, \boldsymbol{x}^t)\, d\gamma(\boldsymbol{x}^s, \boldsymbol{x}^t), \qquad (1)$$

where $\mathcal{X}(\mathcal{D}^s, \mathcal{D}^t)$ denotes the probability distribution between $\mathcal{D}^s$ and $\mathcal{D}^t$. The cost function matrix $\mathcal{M}(\boldsymbol{x}^s, \boldsymbol{x}^t) = ||G_f(\boldsymbol{x}^s) - G_f(\boldsymbol{x}^t)||^k$ denotes the cost to move probability mass from $\boldsymbol{x}^s$ to $\boldsymbol{x}^t$, where $k = 2$ [Damodaran *et al.*, 2018]. The discrete formulation can be expressed as

$$\gamma^* = \underset{\gamma \in \mathcal{X}(\mathcal{D}^s, \mathcal{D}^t)}{\arg\min} \langle \gamma, \mathcal{M} \rangle_F, \qquad (2)$$

where $\gamma^* \in \mathbb{R}^{N \times N}$ is the ideal coupling matrix between the source and target domains, representing as a joint probability measure with source data $\boldsymbol{x}^s$ and target data $\boldsymbol{x}^t$. N is the batchsize. The elements of each row of γ^* are all zeros except one equals 1. $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot product. To separate data of unknown class from known classes in the target

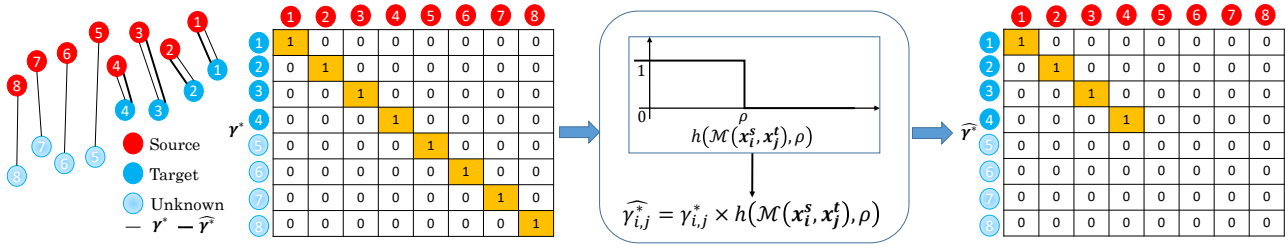


Figure 1: The necessity of the proposed method partial optimal transport for open set domain adaptation. Three colors denote three sets of data. Red: Source samples; Deep Blue: Target common samples. Light Blue: Target unknown samples. The γ^* denotes the optimal transport solution on the global perception which indicates that the unknown samples will be negatively transported to the source domain. Adopting partial optimal transport strategy, exploiting the mean cost ρ of the coupling matching. JDOT chooses the most reliable matching $\hat{\gamma}^*$ for partial domain alignment, eliminating negative transfer caused by unknown class. *Best viewed in color.*

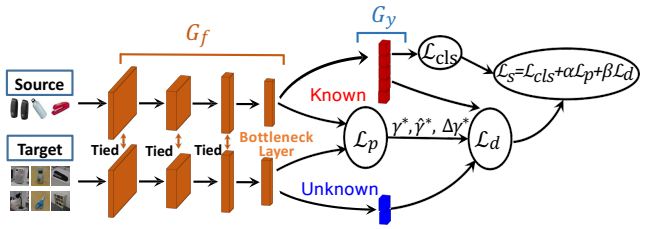


Figure 2: The proposed model **JPOT** for open set domain adaptation. The partial optimal transport loss $\mathcal{L}_p$ is adopted to exploit the unknown target samples then avoid negative transfer. The joint discriminative prototypical compactness loss $\mathcal{L}_d$ utilize the statistics information on the target domain to make features more discriminative. Note that the G_f denotes the CNN network and the fully-connected layers to extract features in the bottleneck layer. G_y denotes the output classification layers. *Best viewed in color.*

domain, we utilize the mean cost of optimal transport matrix ρ to measure the similarity between each target sample and each source sample

$$\rho = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{N}\gamma^*_{i,j}\mathcal{M}(\boldsymbol{x}_i^s, \boldsymbol{x}_j^t) := \frac{1}{N}\sum_{j=1}^{N}\rho_j, \tag{3}$$

where N denotes the number of source samples which equals that of target samples. If the coupling pair distance is bigger than the mean cost ρ, it indicates that this coupling pair is nontransferable. The partial optimal transport for open set domain alignment can be defined as

$$\begin{aligned} \hat{\gamma^*}_{i,j} &= \gamma^*_{i,j} \times h(\mathcal{M}(\boldsymbol{x}_i^s, \boldsymbol{x}_j^t), \rho), \\ h(\mathcal{M}(\boldsymbol{x}_i^s, \boldsymbol{x}_j^t), \rho) &= 1 - \frac{1}{2}(1 + \mathrm{sgn}(\mathcal{M}(\boldsymbol{x}_i^s, \boldsymbol{x}_j^t) - \rho)), \end{aligned} \tag{4}$$

where $\hat{\gamma^*}$ represents partial coupling matrix for known classes of both domains. $sgn(x) = 1$ when $x > 0$, and equals -1 otherwise. Thus, we rank the similarity ρ_j for all target samples and select samples with highest/lowest similarity to align/separate. The loss for known classes in the source and target domains can be defined as

$$\mathcal{L}_o^{kno} = \sum_{i,j}\hat{\gamma^*}_{i,j}(||G_f(\boldsymbol{x}_i^s) - G_f(\boldsymbol{x}_j^t)||^2), \tag{5}$$

which means that the best matched pairs should be pulled each other closely. Meanwhile, with samples "labeled as" unknown class by partial coupling matrix, they should leave away from known classes to avoid negative transfer

$$\mathcal{L}_o^{unk} = \frac{1}{\eta}\sum_{i,j}\Delta\gamma^*_{i,j}\log\left(1 + \exp(-\eta(||G_f(\boldsymbol{x}_i^s) - G_f(\boldsymbol{x}_j^t)||^2))\right), \tag{6}$$

where $\Delta\gamma^*_{i,j} = \gamma^*_{i,j} - \hat{\gamma^*}_{i,j}$ denotes the residual set of the partial optimal transport $\hat{\gamma^*}_{i,j}$. η denotes as a constant and equals to 0.1 in the experiment. Therefore the partial optimal transport loss $\mathcal{L}_p$ is given as

$$\mathcal{L}_p = \mathcal{L}_o^{kno} + \mathcal{L}_o^{unk}. \tag{7}$$

3.2 Joint Discriminative Prototypical Compactness

Inspired by previous work [Herath *et al.*, 2019], exploiting discriminative statistics information on two domains will enhance the transfer performance. When the feature distribution, P_s of domain $\mathcal{D}^s$, is parameterized with the mean $\boldsymbol{\mu}_z^s$ and the covariance $\boldsymbol{\Sigma}_z^s$

$$\begin{aligned} \boldsymbol{\mu}_z^s &= \frac{1}{|N_z|}\sum_{i=1}^{N}G_f(\boldsymbol{x}_i^s)\delta(y_i^s = z), \\ \boldsymbol{\Sigma}_z^s &= \frac{1}{|N_z - 1|}\sum_{i=1}^{N}(G_f(\boldsymbol{x}_i^s) - \boldsymbol{\mu}_z^s)(G_f(\boldsymbol{x}_i^s) - \boldsymbol{\mu}_z^s)^T\delta(y_i^s = z), \end{aligned} \tag{8}$$

where $z = \{1, 2, \cdots, k\}$ denotes the corresponding shared known classes, k denotes the number of classes in $\mathcal{D}^s$, $\delta(x) = 1$ when x is True, and $\delta(x) = 0$ when x is False. The variance vanishes when there is only one sample in the batch. Due to the lack of reliable labels in the target domain, it is difficult to measure the unknown samples in the target domain distribution P_t. The most difficult task is to find the mean and variance of unknown samples. According to the partial optimal transport, we provide unknown index U_j to measure the probability of sample x_j^t to be in the Unknown class.

$$U_j = \sum_{i=1}^{N}\Delta\gamma^*_{i,j}, \tag{9}$$

The number of the unknown class samples can be given as

$$N_{unk}^t = \sum_{j=1}^{N}U_j. \tag{10}$$

The mean $\boldsymbol{\mu}_{unk}$ and the covariance $\boldsymbol{\Sigma}_{unk}$ can be described as

$$\boldsymbol{\mu}_{unk}^{t} = \frac{1}{N_{unk}^{t}} \sum_{j=1}^{N} G_f(\boldsymbol{x}_j^t) \times U_j, \tag{11}$$

$$\boldsymbol{\Sigma}_{unk}^{t} = \frac{1}{|N_{unk}^{t} - 1|} \sum_{j=1}^{N} U_j \times (G_f(\boldsymbol{x}_j^t) - \boldsymbol{\mu}_{unk}^{t})(G_f(\boldsymbol{x}_j^t) - \boldsymbol{\mu}_{unk}^{t})^T. \tag{12}$$

To quantify spatial prototypical information in both domains, we define $\boldsymbol{\mu}^A = \{\boldsymbol{\mu}_1^s, \boldsymbol{\mu}_2^s, \cdots, \boldsymbol{\mu}_k^s, \boldsymbol{\mu}_{unk}^t\}$ and $\boldsymbol{\Sigma}^A = \{\boldsymbol{\Sigma}_1^s, \boldsymbol{\Sigma}_2^s, \cdots, \boldsymbol{\Sigma}_k^s, \boldsymbol{\Sigma}_{unk}^t\}$ for computation. Therefore $\boldsymbol{\mu}^A \in \mathbb{R}^{(k+1)\times d}$ and $\boldsymbol{\Sigma}^A \in \mathbb{R}^{(k+1)\times d\times d}$. d is the number of output neurons in the bottleneck layer. The spatial prototypical matrix $\mathbf{D} \in \mathbb{R}^{(k+1)\times N}$ is defined as

$$D(z, j) = \frac{e^{-d(G_f(\boldsymbol{x}_j^t), \boldsymbol{\mu}_z^A)}}{\sum_{m=1}^{k+1} e^{-d(G_f(\boldsymbol{x}_j^t), \boldsymbol{\mu}_m^A)}}, \tag{13}$$

where $d(G_f(\boldsymbol{x}_j^t), \boldsymbol{\mu}_z^A)$ is the distance between the target sample $G_f(\boldsymbol{x}_j^t)$ and z-th class center $\boldsymbol{\mu}_z^A$. Therefore, the Mahalanobis distance formulation of $d(G_f(\boldsymbol{x}_j^t), \boldsymbol{\mu}_z^A)$ can be defined as

$$d(G_f(\boldsymbol{x}_j^t), \boldsymbol{\mu}_z^A) = \sqrt{(G_f(\boldsymbol{x}_j^t) - \boldsymbol{\mu}_z^A)^T (\boldsymbol{\Sigma}_z^A)^{-1} (G_f(\boldsymbol{x}_j^t) - \boldsymbol{\mu}_z^A)}. \tag{14}$$

The joint discriminative prototypical compactness loss $\mathcal{L}_d$ is defined as

$$\mathcal{L}_d = \mathcal{L}_{dc} + \mathcal{L}_{dp}, \tag{15}$$

The $\mathcal{L}_{dc}$ denotes the centroid clustering loss in the source class domain

$$\mathcal{L}_{dc} = \sum_{i=1}^{N} \left|\left| G_f(\boldsymbol{x}_i^s) - \boldsymbol{\mu}_{y_i^s}^s \right|\right|_2^2. \tag{16}$$

The $\mathcal{L}_{dp}$ denotes the prototypical compactness loss for the target samples

$$\mathcal{L}_{dp} = \sum_{z=1}^{k+1} \sum_{j=1}^{N} D(z, j) \mathcal{F}(Softmax(G_y(G_f(\boldsymbol{x}_j^t)), z)), \tag{17}$$

where $\mathcal{F}(\cdot, \cdot)$ denotes the cross-entropy loss. Furthermore, we introduce an additional loss $\mathcal{L}_{cls}$ to minimize the probability of source samples to be misclassified as unknown class.

$$\mathcal{L}_{cls} = \sum_{i=1}^{N} \mathcal{F}(Softmax(G_y(G_f(\boldsymbol{x}_i^s), k+1)). \tag{18}$$

In conclusion, the total loss of our JPOT model is

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha \mathcal{L}_p + \beta \mathcal{L}_d, \tag{19}$$

where the α and β denote the trade-off parameters.

To calculate the statistic information, we adopt the incremental learning strategy as follows:

$$\begin{aligned} \boldsymbol{\mu}_{accu.}^{A(t)} &= m \times \boldsymbol{\mu}_{accu.}^{A(t-1)} + (1-m) \times \boldsymbol{\mu}^{A(t)}, \\ \boldsymbol{\Sigma}_{accu.}^{A(t)} &= m \times \boldsymbol{\Sigma}_{accu.}^{A(t-1)} + (1-m) \times \boldsymbol{\Sigma}^{A(t)}, \end{aligned} \tag{20}$$

where $\boldsymbol{\mu}_{accu.}^{A(t)}$ and $\boldsymbol{\Sigma}_{accu.}^{A(t)}$ denote the accumulation results of the mean and variance, $0 \leq m \leq 1$ is the momentum hyperparameter for the accumulation.

4 Experiments

Now we evaluate our method with state-of-the-art domain adaptation approaches on several benchmark datasets.

4.1 Setup

Digits contains three standard digit classification datasets: *MNIST* [Lecun *et al.*, 1998], *USPS* [Hull, 2002], and *SVHN* [Netzer *et al.*, 2011]. Each dataset consists of 10 classes of digits, ranging from 0 to 9. Following the same evaluation protocol of [Saito *et al.*, 2018], we construct three open set domain adaptation tasks and report adaptation results on the test sets: **SVHN→MNIST** (one task) and **MNIST↔USPS** (two tasks).

Office-31 [Saenko *et al.*, 2010] is a standard dataset in computer vision for domain adaptation which contains 4652 images from 31 categories with three domains: *Amazon* (**A**), *Webcam* (**W**) and *DSLR* (**D**). According to previous work [Liu *et al.*, 2019], we adopt the same set of known classes and unknown classes in the target domain and evaluate all methods on following four challenging settings: **A↔W** and **A↔D**.

Office-Home [Venkateswara *et al.*, 2017] is a more challenging domain adaptation dataset which consists of around 15500 images from 65 object classes in 4 distinct domains: *Aetistic* (**Ar**), *Clipart* (**Cl**), *Product* (**Pr**), and *Real-Word* (**Rw**). We select (in alphabetic order) the first 25 classes as known classes shared by the source and target domains [Saito *et al.*, 2018]. The 26-65 classes are regarded as the unknown class. We report performances of all the 12 adaptation tasks to enable thorough evaluations: **Ar↔Cl**, **Ar↔Pr**, **Ar↔Rw**, **Cl↔Pr**, **Cl↔Rw**, and **Pr↔Rw**.

Compared Approaches. We mainly compare our proposal with several open set recognition, domain adaptation and open set domain adaptation methods as previous work [Liu *et al.*, 2019]: Open Set SVM(**OSVM**) [Jain *et al.*, 2014], **DANN** [Ganin and Lempitsky, 2015], **RTN** [Long *et al.*, 2016], **OpenMAX** [Bendale and Boult, 2016], **MMD+OSVM**, **DANN+OSVM**, **OSBP** [Saito *et al.*, 2018] and **STA** [Liu *et al.*, 2019]. MMD+OSVM and DANN+OSVM are two variants of OSVM that incorporate Maximum Mean Discrepancy [Tzeng *et al.*, 2014] and domain adversarial network [Ganin and Lempitsky, 2015] in OSVM. In our experiments, we compare the average accuracy of each method on five random experiments.

Evaluation Metrics. Following previous works [Saito *et al.*, 2018], we employ four evaluation metrics for open set domain adaptation: **OS:** the accuracy averaged over all the classes including the unknown class as one class; **OS*:** the accuracy averaged only on known classes; **ALL:** the accuracy of all samples; **UNK:** the accuracy of unknown instances.

4.2 Implementation Details

For the experiments on *Office-31* and *Office-Home*, our proposal and the compared approaches are both trained by finetuning with ResNet-50 [He *et al.*, 2016] pre-trained on ImageNet. And we adopt LeNet [Lecun *et al.*, 1998] to investigate the efficacy of our framework for the experiments on *Digits* datasets. We use all labeled source examples and unlabeled examples for training.

Method	A→W		A→D		D→W		W→D		D→A		W→A		Avg	
	OS	OS*	OS	OS*	OS	OS*	OS	OS*	OS	OS*	OS	OS*	OS	OS*
ResNet	82.5±1.2	82.7±0.9	85.2±0.3	85.5±0.9	94.1±0.3	94.3±0.7	96.6±0.2	97.0±0.4	71.6±1.0	71.5±1.1	75.5±1.0	75.2±1.6	84.2	84.4
RTN	85.6±1.2	88.1±1.0	89.5±1.4	90.1±1.6	94.8±0.3	96.2±0.7	97.1±0.2	98.7±0.9	72.3±0.9	72.8±1.5	73.5±0.6	73.9±1.4	85.4	86.8
DANN	85.3±0.7	87.7±1.1	86.5±0.6	87.7±0.6	97.5±0.2	**98.3**±0.5	**99.5**±0.1	**100.0**±.0	75.7±1.6	76.2±0.9	74.9±1.2	75.6±0.8	86.6	87.6
OpenMax	87.4±0.5	87.5±0.3	87.1±0.9	88.4±0.9	96.1±0.4	96.2±0.3	98.4±0.3	98.5±0.3	83.4±1.0	82.1±0.6	82.8±0.9	82.8±0.6	89.0	89.3
OSBP	86.5±2.0	87.6±2.1	88.6±1.4	89.2±1.3	97.0±1.0	96.5±0.4	97.9±0.9	98.7±0.6	88.9±2.5	90.6±2.3	85.8±2.5	84.9±1.3	90.8	91.3
STA	89.5±0.6	92.1±0.5	93.7±1.5	**96.1**±0.4	97.5±0.2	96.5±0.5	**99.5**±0.2	99.6±0.1	89.1±0.5	93.5±0.8	87.9±0.9	87.4±0.6	92.9	94.1
DeepJDOT	86.1±0.5	88.7±0.9	86.9±0.7	89.0±0.5	96.9±0.2	95.6±0.8	96.1±0.1	98.1±0.5	85.6±1.2	88.4±1.0	81.5±1.3	83.1±0.9	88.9	90.5
JPOT	**92.8**±0.6	**92.2**±0.4	**95.2**±0.9	96.0±0.6	**98.1**±0.3	96.2±0.4	**99.5**±0.1	98.6±0.2	**93.0**±0.7	**94.1**±0.4	**88.9**±1.0	**88.4**±0.4	**94.6**	**94.3**

Table 1: Classification accuracy (%) on *Office-31* for open set domain adaptation (ResNet-50).

Method	Ar→Cl	Pr→Cl	Rw→Cl	Ar→Pr	Cl→Pr	Rw→Pr	Cl→Ar	Pr→Ar	Rw→Ar	Ar→Rw	Cl→Rw	Pr→Rw	Avg
ResNet	53.4±0.4	52.7±0.6	51.9±0.5	69.3±0.7	61.8±0.5	74.1±0.4	61.4±0.6	64.0±0.3	70.0±0.3	78.7±0.6	71.0±0.6	74.9±0.9	65.3
DANN	54.6±0.7	49.7±1.6	51.9±1.4	69.5±1.1	63.5±1.0	72.9±0.8	61.9±1.2	63.3±1.0	71.3±1.0	80.2±0.8	71.7±0.4	74.2±0.4	65.4
OpenMax	56.5±0.4	52.9±0.7	53.7±0.4	69.1±0.3	64.8±0.4	74.5±0.6	**64.1**±0.9	64.0±0.8	71.2±0.8	80.3±0.8	73.0±0.5	76.9±0.3	66.7
OSBP	56.7±1.9	51.5±2.1	49.2±2.4	67.5±1.5	65.5±1.5	74.0±1.5	62.5±2.0	64.8±1.1	69.3±1.1	80.6±0.9	74.7±2.2	71.5±1.9	65.7
STA	58.1±0.6	53.1±0.9	54.4±1.0	71.6±1.2	69.3±1.0	81.9±0.5	63.4±0.5	65.2±0.8	74.9±1.0	**85.0**±0.2	75.8±0.4	80.8±0.3	69.5
DeepJDOT	56.7±0.8	50.4±1.1	53.7±1.2	67.1±1.4	64.4±0.8	76.2±0.7	62.5±0.6	64.9±1.2	72.5±1.0	82.1±0.7	74.0±0.5	77.1±0.6	66.8
JPOT	**59.6**±0.5	**54.2**±0.7	**54.6**±0.9	**72.3**±1.1	**70.1**±0.6	**82.1**±0.9	62.9±0.7	**68.3**±0.8	**75.1**±1.1	84.8±0.4	**77.4**±0.5	**81.2**±0.4	**70.2**

Table 2: Classification accuracy OS (%) on *Office-Home* for open set domain adaptation (ResNet-50).

Method	SVHN→MNIST				USPS→MNIST				MNIST→USPS				Avg			
	OS	OS*	ALL	UNK	OS	OS*	ALL	UNK	OS	OS*	ALL	UNK	OS	OS*	ALL	UNK
OSVM	54.3	63.1	37.4	10.5	43.1	32.3	63.5	97.5	79.8	77.9	84.2	89.0	59.1	57.7	61.7	65.7
MMD+OSVM	55.9	64.7	39.1	12.2	62.8	58.9	69.5	82.1	80.0	79.8	81.3	81.0	68.0	68.8	66.3	58.4
DANN+OSVM	62.9	75.3	39.2	0.70	84.4	92.4	72.9	0.90	33.8	40.5	21.4	44.3	60.4	69.4	44.5	15.3
OSBP	63.0	59.1	71.0	82.3	**92.3**	91.2	**94.4**	97.6	92.1	**94.9**	88.1	78.0	82.4	81.7	84.5	85.9
STA	76.9	**75.4**	80.0	84.4	92.2	**91.3**	93.9	96.5	**93.0**	**94.9**	90.3	83.5	87.3	**87.2**	88.1	88.1
DeepJDOT	61.2	71.6	41.3	9.21	85.6	89.4	79.4	66.7	83.9	85.8	87.9	74.4	76.9	82.3	69.5	50.1
JPOT	**79.2**	75.3	**80.8**	**86.7**	**92.4**	91.2	**94.4**	**98.4**	92.9	92.1	**93.9**	**96.9**	**88.2**	85.4	**89.7**	**94.0**

Table 3: Classification accuracy (%) on *Digits* for open set domain adaptation (LeNet).

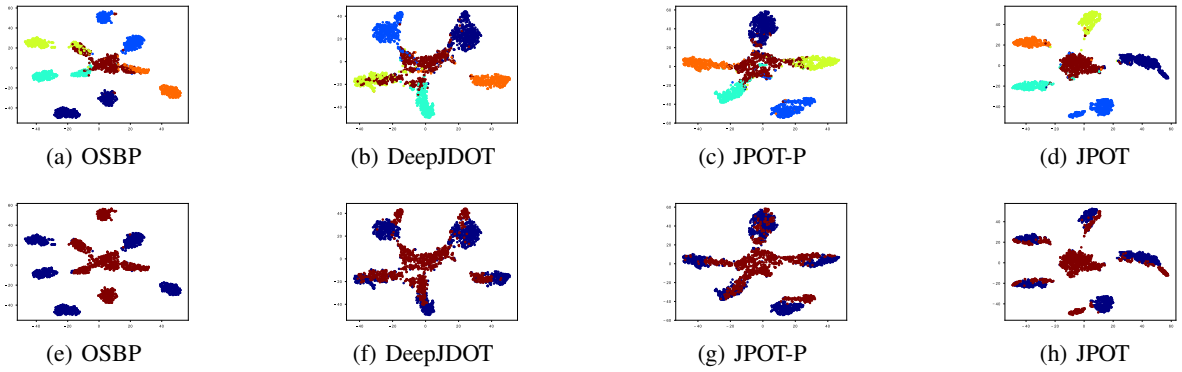
All the mentioned deep learning methods are trained with Adam optimizer. And the model is trained on 256-sized batches totally with $N = 128$ samples from each domain. We use mini-batch SGD with momentum to 0.9 and the same learning rate strategy in [Saito *et al.*, 2018]. Following previous work [Herath *et al.*, 2019], m is set as 0.4 in the whole experiment. Note that the parameter η is set as 0.1. As for tradeoff hyper-parameters α and β, we select $\alpha = 0.02$ and $\beta = 0.05$ for all transfer tasks.

4.3 Result and Discussion

The classification accuracy results on the *Office-31* dataset for open set domain adaptation based on ResNet-50 are shown in Table 1. For fair comparison, all results of compared approaches are directly reported from their original papers. The JPOT exceeds the performance of all previous methods on most transfer tasks. It is worth noting that our proposal improves the classification accuracies substantially on hard transfer tasks, e.g. **A**→**W** and **W**→**A**, and achieves comparable classification performance on easy transfer tasks, e.g. **D**→**W** and **W**→**D**. In particular, JPOT produces higher **OS** than **OS***, verifying that JPOT is robust to open set domain adaptation scenarios. Besides, some close set domain adaptation methods perform worse than ResNet on a few tasks since these methods suffer from negative transfer caused by mismatching unknown classes in the target domain to known classes in the source domain. JPOT aligns the discriminative representations of shared class centers in both source and target domains by center-based optimal transport strategy, and achieves intra-class compactness and inter-class separability.

Results on the twelve tasks of *Office-Home* dataset are shown in Table 2. Due to the large domain gap between the source and target domains, we observe that previous open set domain adaptation approaches obtain poor performance on some tasks. The JPOT approach outperforms the comparison methods on most transfer tasks. The encouraging results suggest that our proposal alleviates the negative transfer issue brought by the unknown class in the target domain, and also highlight the importance of center-based optimal transport strategy in open set domain adaptation.

We further compare JPOT with previous approaches on the *Digits* dataset, as reported in Table 3. In contrast to *Office-31* dataset, *Digits* dataset has a much larger domain size. JPOT overpasses all comparison methods with different metrics on most transfer tasks. Note that, **SVHN** dataset contains significant variations (in scale, slanting, blurring and rotation) and confounding information from background that makes large domain gap on task **SVHN**→**MNIST**. JPOT achieves better performance than STA on task **SVHN**→**MNIST**. In particular, JPOT almost achieves the state-of-the-art performance with the **UNK** accuracy in all transfer tasks, which suggests



(a) OSBP (b) DeepJDOT (c) JPOT-P (d) JPOT

(e) OSBP (f) DeepJDOT (g) JPOT-P (h) JPOT

Figure 3: Visualization of features on MNIST→USPS task. Figure (a-d) represents category alignment (Each color denotes a class. The Red denotes the Unknown class). Figure (e-h) represents domain alignment (Blue: Source domain; Red: Target domain). *Best viewed in color.*

Method	MNIST→USPS			
	OS	OS*	ALL	UNK
JPOT-P	92.0	91.1	89.2	90.4
JPOT-T	91.4	91.1	87.9	87.4
JPOT-S	92.4	91.7	92.5	93.1
JPOT	92.9	92.1	93.9	96.9

Table 4: Ablation study: classification accuracy (%) on *Digits*.

that JPOT is robust to large domain gap and able to learn more transferable representations for open set domain adaptation.

4.4 Ablation Study

To tooth apart the separate contributions of center-based optimal transport strategy and discriminative domain alignment, we compare JPOT with OSBP [Saito *et al.*, 2018], DeepJDOT [Damodaran *et al.*, 2018] and a variant of JPOT using the t-SNE embeddings [Donahue *et al.*, 2014] of the last-layer features on transfer task **MNIST→USPS** in Figures 3(a)-3(h). (**1**) In Figures 3(a) and 3(e), we observe that in the OSBP model, scattered samples of unknown and known classes in the target domain are close or even mixed together and the domains are not well aligned, which may cause negative transfer. Therefore the OSBP output classification is of low confidence. (**2**) In Figures 3(b) and 3(f), although the original DeepJDOT computes the global optimal transport solution, the target unknown samples still be mistaken transported to the corresponding source domain. Apparently the unknown samples mixes up with the known samples which leads to terrible prediction results. (**3**) As shown in Figures 3(c) and 3(g), compared with OSBP and DeepJDOT, **JPOT-P** ($\mathcal{L}_{cls} + \alpha\mathcal{L}_p$) with partial optimal transport strategy and without joint discriminative prototypical compactness loss, achieves better domain alignment with high classification performance meanwhile avoid negative transfer. Due to the fact that partial optimal transport exploits the most likely to be the unknown samples, the unknown target samples will not fully mix up with the known samples. (**4**) As shown in Figures 3(d) and 3(h), JPOT achieves intra-class compactness and inter-class separability, and also improves with a large room over JPOT-P, validating the complement of discriminative feature learning with the corporation of partial optimal transport.

In Table 4, further ablation tests have been carried out on variations of JPOTs including **JPOT-T** ($\mathcal{L}_{cls} + \alpha\mathcal{L}_o^{kno}$), **JPOT-S** ($\mathcal{L}_{cls}+\alpha\mathcal{L}_p+\beta\mathcal{L}_{dc}$), and **JPOT-P**. (1) While **JPOT-T** still works well on detecting unknown instances (87.4% of accuracy), the performance falls behind other JPOT variations. Without the additional loss $\mathcal{L}_o^{unk}$, **JPOT-T** lacks driving force to push unknown samples away from the known samples in the feature space, resulted in inferior performance. (2) **JPOT-S** outperforms **JPOT-P** on all metrics: this indicates that joint discriminative learning especially the compactifying effect due to L_{dc} to make source samples more close to their class means could help to reduce the misclassification error caused by those source samples scattered near the the decision boundary. (3) JPOT outperforms all variations. This demonstrates that the prototypical compactness loss $\mathcal{L}_{dp}$ to compactify target samples and to be more discriminative is a good strategy to boost the efficacy significantly.

5 Conclusion

This paper presents Joint Partial Optimal Transport (JPOT) for open set domain adaptation, addressing the critical challenge of negative transfer brought by unknown class samples in the target domain. We propose partial optimal transport based on global optimal transport with coupling matching matrix to exploit the most likely unknown target samples. Meanwhile, the joint discriminative feature learning helps to not only match shared class samples in both domains, but also separate samples of unknown and known classes in the target domain. This well improves the domain adaptation performance. Comprehensive experiments show that our method outperforms state-of-the-art results on various domain adaptation datasets.

Acknowledgements

This research was supported by National Key Research and Development Program of China (No. 2017YFB1400603), National Natural Science Foundation of China under grants (No. 61825205, No. 61772459, No. 61402403), and MoE Engineering Research Center of Digital Library.

References

[Angenent *et al.*, 2003] S. Angenent, S. Haker, and A. Tannenbaum. Minimizing flows for the monge–kantorovich problem. *SIAM journal on mathematical analysis*, 35(1):61–97, 2003.

[Bendale and Boult, 2016] A. Bendale and T.E. Boult. Towards open set deep networks. In *Proceedings CVPR*, pages 1563–1572, 2016.

[Chen *et al.*, 2019] C. Chen, Z. Chen, B. Jiang, and X. Jin. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. In *Proceedings AAAI*, volume 33, pages 3296–3303, 2019.

[Courty *et al.*, 2016] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[Courty *et al.*, 2017] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3730–3739, 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*, pages 467–483, 2018.

[Donahue *et al.*, 2014] J. Donahue, Yangqing Jia, O. Vinyals, J. Hoffman, Ning Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[Ganin and Lempitsky, 2015] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015.

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

[Herath *et al.*, 2019] S. Herath, M. Harandi, B. Fernando, and R. Nock. Min-max statistical alignment for transfer learning. In *Proceedings CVPR*, pages 9288–9297, 2019.

[Hull, 2002] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 16(5):550–554, 2002.

[Jain *et al.*, 2014] L.P. Jain, W.J. Scheirer, and T.E. Boult. Multi-class open set recognition using probability of inclusion. In *ECCV*, pages 393–409. Springer, 2014.

[Kang *et al.*, 2019] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings CVPR*, pages 4893–4902, 2019.

[Lecun *et al.*, 1998] Y. L. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings IEEE*, 86(11):2278–2324, 1998.

[Liu *et al.*, 2019] H. Liu, Z. Cao, M. Long, J. Wang, and Q. Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings CVPR*, pages 2927–2936, 2019.

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

[Long *et al.*, 2016] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *NeurIPS*, pages 136–144, 2016.

[Luo *et al.*, 2017] Z. Luo, Y. Zou, J. Hoffman, and F. Li. Label efficient learning of transferable representations across domains and tasks. In *Advances in Neural Information Processing Systems*, pages 165–177, 2017.

[Netzer *et al.*, 2011] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. *Nips Workshop on Deep Learning & Unsupervised Feature Learning*, 2011.

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

[Pan *et al.*, 2019] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings CVPR*, pages 2239–2247, 2019.

[S. and S., 2016] Baochen S. and Kate S. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshops*, 2016.

[Saenko *et al.*, 2010] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[Saito *et al.*, 2018] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada. Open set domain adaptation by backpropagation. In *Proceedings ECCV*, pages 153–168, 2018.

[Snell *et al.*, 2017] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4077–4087, 2017.

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

[Venkateswara *et al.*, 2017] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings CVPR*, pages 5018–5027, 2017.

[Yan *et al.*, 2018] Y. Yan, W. Li, H. Wu, H. Min, M. Tan, and Q. Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *IJCAI*, pages 2969–2975, 2018.

[You *et al.*, 2019] K. You, M. Long, Z. Cao, J. Wang, and M. I Jordan. Universal domain adaptation. In *Proceedings CVPR*, pages 2720–2729, 2019.

[Zellinger *et al.*, 2017] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv:1702.08811*, 2017.