Metadata of the article that will be visualized in OnlineFirst

1	Article Title	**Class-rebalanced wasserstein distance for multi-source domain adaptation**	
2	Article Sub-Title		
3	Article Copyright - Year	**Springer Science+Business Media, LLC, part of Springer Nature 2022** **(This will be the copyright line in the final PDF)**	
4	Journal Name	Applied Intelligence	
5	Corresponding Author	Family Name	**Wang**
6		Particle	
7		Given Name	**Shengsheng**
8		Suffix	
9		Organization	Jilin University
10		Division	College of Software
11		Address	Changchun, 130012, Jilin, China
12		Organization	Jilin University
13		Division	Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education
14		Address	Changchun, 130012, Jilin, China
15		Organization	Jilin University
16		Division	College of Computer Science and Technology
17		Address	Changchun, 130012, Jilin, China
18		e-mail	wss@jlu.edu.cn
19	Author	Family Name	**Wang**
20		Particle	
21		Given Name	**Qi**
22		Suffix	
23		Organization	Jilin University
24		Division	College of Software
25		Address	Changchun, 130012, Jilin, China
26		Organization	Jilin University
27		Division	Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education
28		Address	Changchun, 130012, Jilin, China
29		e-mail	wangqi20@mails.jlu.edu.cn

Author	Family Name	**Wang**
	Particle	
	Given Name	**Bilin**
	Suffix	
	Organization	Jilin University
	Division	Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education
	Address	Changchun, 130012, Jilin, China
	Organization	Jilin University
	Division	College of Computer Science and Technology
	Address	Changchun, 130012, Jilin, China
	e-mail	blwang19@mails.jlu.edu.cn

Schedule	Received	
	Revised	
	Accepted	23 May 2022

Abstract

In the study of machine learning, multi-source domain adaptation (MSDA) handles multiple datasets which are collected from different distributions by using domain-invariant knowledge extraction. However, the current studies mainly employ features and raw labels on the joint space to perform domain alignment, neglecting the intrinsic structure of label distribution that can harm the performance of adaptation. Therefore, to make better use of label information when aligning joint feature-label distribution, we propose a rebalancing scheme, class-rebalanced Wasserstein distance (CRWD), for unsupervised MSDA under class-wise imbalance and data correlation. Based on the optimal transport for domain adaptation (OTDA) framework, CRWD mitigates the impact of the biased label structure by rectifying the Wasserstein mapping from source to target space. Technically, the class proportions are utilized to encourage distributional transportation between minor classes and principal components, which reweigh the optimal transport plan and reinforce the ground metric of Mahalanobis distance to better metricise the differences among domains. In addition, the scheme measures both inter-domain and intra-source discrepancies to enhance adaptation. Extensive experiments are conducted on various benchmarks, and the results prove that CRWD has competitive advantages.

Keywords separated by ' - ': Domain adaptation - Data correlation - Class imbalance - Multiple sources - Optimal transport

Foot note information

Applied Intelligence
https://doi.org/10.1007/s10489-022-03810-y

Class-rebalanced wasserstein distance for multi-source domain adaptation

Q1 Qi Wang[1,2] · Shengsheng Wang[1,2,3] · Bilin Wang[2,3]

Accepted: 23 May 2022

Abstract

In the study of machine learning, multi-source domain adaptation (MSDA) handles multiple datasets which are collected from different distributions by using domain-invariant knowledge extraction. However, the current studies mainly employ features and raw labels on the joint space to perform domain alignment, neglecting the intrinsic structure of label distribution that can harm the performance of adaptation. Therefore, to make better use of label information when aligning joint feature-label distribution, we propose a rebalancing scheme, class-rebalanced Wasserstein distance (CRWD), for unsupervised MSDA under class-wise imbalance and data correlation. Based on the optimal transport for domain adaptation (OTDA) framework, CRWD mitigates the impact of the biased label structure by rectifying the Wasserstein mapping from source to target space. Technically, the class proportions are utilized to encourage distributional transportation between minor classes and principal components, which reweigh the optimal transport plan and reinforce the ground metric of Mahalanobis distance to better metricise the differences among domains. In addition, the scheme measures both inter-domain and intra-source discrepancies to enhance adaptation. Extensive experiments are conducted on various benchmarks, and the results prove that CRWD has competitive advantages.

Keywords Domain adaptation · Data correlation · Class imbalance · Multiple sources · Optimal transport

1 Introduction

In the science of artificial intelligence, machine learning has exhibited a promising capability of simulating and anticipating complex scenarios. However, it is time-consuming and unaffordable to manually annotate a large amount of unlabeled data. In addition, conventional machine learning paradigms, namely, supervised, semi-supervised, and unsupervised learning, assume that data collected from various sources obey the same probability distribution [28, 37]. The premise is not guaranteed in many real-world environments due to settings such as data format, background noise, and lighting conditions [16, 17]. Ignoring such differences by directly training a model with source data and then applying it to the target will result in poor generalization, which is known as *negative transfer* [48]. Therefore, aimed at mitigating such data shifts, domain adaptation (DA), which is a subfield of transfer learning, applies shared knowledge extracted from labeled information (or source domain D_S) to gain better performance on unlabeled data (or target domain D_T). A typical adaptation is displayed in Fig. 1.

Based on the quantity of source domains, DA can be classified into single-source domain adaptation (SSDA) or multi-source domain adaptation. Notably, the setting of MSDA is particularly challenging as it is defined to extract domain-invariant knowledge from more comprehensive sources, where the comprehensiveness can be caused by the increasing quantity of source domains and the inner structure of datasets. Technically, once domain-specific

✉ Shengsheng Wang
wss@jlu.edu.cn

Qi Wang
wangqi20@mails.jlu.edu.cn

Bilin Wang
blwang19@mails.jlu.edu.cn

[1] College of Software, Jilin University, Changchun, 130012, Jilin, China

[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012, Jilin, China

[3] College of Computer Science and Technology, Jilin University, Changchun, 130012, Jilin, China

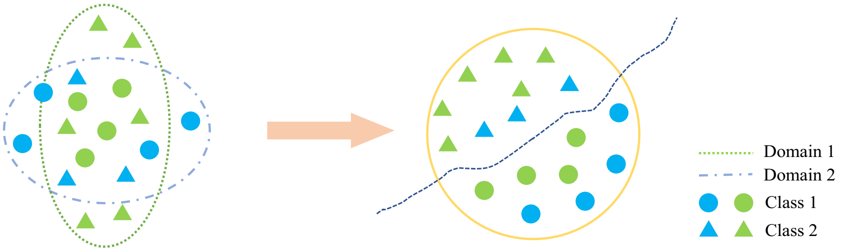


Fig. 1 A typical domain adaptation process. The left side shows that before adaptation, the blue and green shapes are subject to different distributions, and it is challenging to draw a boundary between different classes. The right side reveals that after extracting domain-invariant knowledge and transforming data into the adapted space with the knowledge, the boundary can be easily decided (Best viewed in color)

knowledge is screened out by certain measurement, the remaining common features, representing the intersection of multiple different domains, can be adopted for inferring labels on target data. In order to generate such features, most MSDA solutions depend on two paradigms: **i**) Plain feature-level alignment. For instance, the researchers [26, 43] deploy the alignment process on generative adversarial networks (GAN) to generate common features using a minimax game. However, this kind of method ignores that labels can also carry domain information [2, 36], which contributes to achieving optimality; **ii**) Performing domain alignment on the joint feature-label distribution. Our work follows the second paradigm. Several studies [7, 40] indeed attempt to implement label information into designed neural networks. Nevertheless, their methods neglect that the comprehensiveness of sources, such as the biased structure of datasets, can still hamper models from being robust and generative. Actually, due to the increasing quantities of domain and manners of acquisition, the proportion of classes over all domains can be highly biased, which is known as *label shift* $P_S(\mathbf{y}) \neq P_T(\mathbf{y})$, causing conventional methods to behave overfitting on major classes and underfitting on minor classes. Moreover, some samples labeled as one class may contain features mixed with another, which indicates that features of different categories are correlated to certain extent (see Fig. 3). Such data correlation rooted in multiple domains can contribute to *conditional shift*, where the conditional distribution $P(\mathbf{x} \mid \mathbf{y})$ is inconsistent across domain, i.e., $P_S(\mathbf{x} \mid \mathbf{y}) \neq P_T(\mathbf{x} \mid \mathbf{y})$. According to the Bayes' theorem [19, 35]:

$$P_S(\mathbf{y} \mid \mathbf{x}) = \frac{P_s(\mathbf{x} \mid \mathbf{y}) P_s(\mathbf{y})}{P_s(\mathbf{x})}$$
$$P_T(\mathbf{y} \mid \mathbf{x}) = \frac{P_T(\mathbf{x} \mid \mathbf{y}) P_T(\mathbf{y})}{P_T(\mathbf{x})} \quad (58)$$

such biased intrinsic structure of class imbalance and data correlation violates the assumption which guarantees the existence of the optimal classifier that is consistent across domain, i.e., $P_S(\mathbf{y} \mid \mathbf{x}) = P_T(\mathbf{y} \mid \mathbf{x})$. Obviously, it can be seen from the theorem that class information equally contributes to the performance of adaptation. Henceforth, our proposed model is motivated to exploit the inner structure of label distribution and to alleviate its impact.

In this study, we propose CRWD to address the unsupervised MSDA problem under unbalanced class and correlated data distribution. As presented in Fig. 2, our model consists of three components including a generator component that extracts raw features from input data, followed by a data alignment component that jointly maps feature and labels information into a common latent space. This stage is based on the powerful Wasserstein distance (or OT distance) to exploit class proportions with the purpose of gaining unbiased transformation from source data space to target one. In the case of correlated data, the proportions are also harnessed for discriminative feature enhancement. In the end, the transformed source data is fed into the classification component, which gives prediction on label space. Compared with previous works, our method introduces fewer parameters than GAN-based models, and can rectify the impact of class disequilibrium and data correlation for obtaining better benefit from label information.

We present the main contributions of this study as follows:

(1) The solution constructs a class-rebalanced Wasserstein space, which alleviates the impact of unbalanced samples of minor and major classes, and metricizes the difference between multiple domains more precisely.
(2) The class-rebalancing scheme is also leveraged to enhance the distributional transportation between principal components, improving the performance on highly correlated data.
(3) Extensive experiments are conducted to demonstrate the effectiveness of the proposed model.

The rest of this paper is organized as follows: Section 2 presents a review of related works in the field of MSDA and OT theory. The notation used in this study, problem definition for MSDA and proposed method of CRWD are detailed in Section 3. Section 4 provides information concerning the selected datasets and implementation settings of the proposed networks, together with the results and

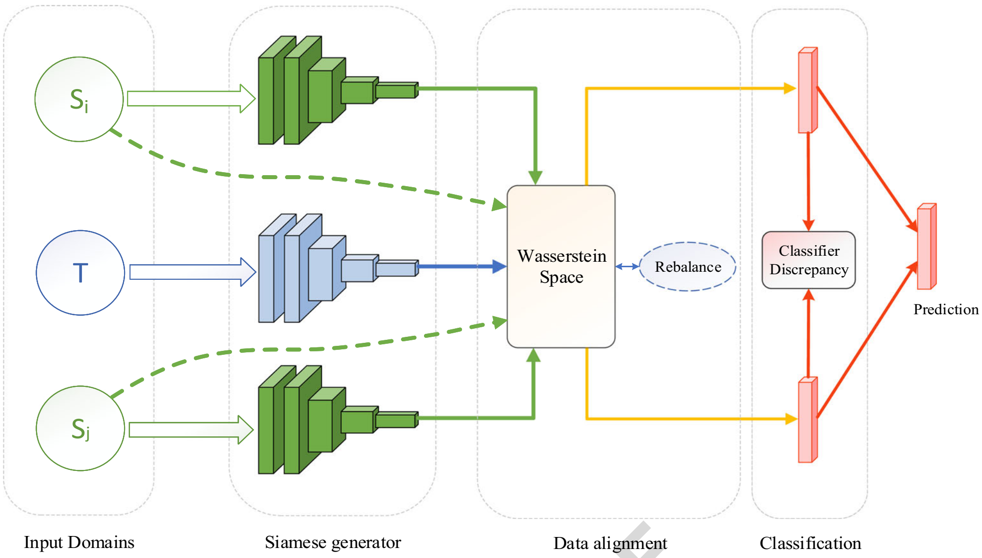


Fig. 2 The framework of CRWD. The green dashed lines stand for label information. The solid lines in green, blue and orange indicate features before and after adaptation through the rectified Wasserstein space. Obviously, the weights of neurons among Siamese generators are shared in the generator component, which is equivalent to a single generator. Here, blue and green colors for generators are only used for domain identification

analysis. Finally, we draw the conclusion and discuss the future works in Section 5.

2 Related work

2.1 Multi-source domain adaptation

Unsupervised MSDA focuses on utilizing knowledge that is harvested from well-annotated source domains to achieve better performance on fully-unlabeled target one. By performing the survey [32], most MSDA methods can be categorized into two different learning strategies. Based on the theoretical analysis by Mansour et al. [20], instance reweighting estimates the optimal target distribution by learning a mixture of selected source samples. For example, [4, 13] lowered the marginal distribution difference (or covariate shift) between the source and target domains by training a support vector machine (SVM) classifier, and the statistical risks on predicted labels are treated as domain weights. [38] developed an attention scheme together with conditional Wasserstein distance between domains as corresponding values to reweigh multiple sources. However, this kind of method depends on the hypothesis [20] that both source and target domains share an identical conditional distribution, which can be violated with regard to the pattern of real-world data generation.

Another extensively proposed approach refers to common feature extraction, which aims to exploit domain-invariant features by minimizing specific discrepancies between empirical data distributions. Therefore, the extracted shared features can be utilized to train a generative model for all domains. The proposed method relies heavily on the discriminative ability of the designed discrepancy, and several studies attempt to exploit this metric by using maximum mean discrepancy (MMD) [14, 39], second-order statistics, *i.e.* correlation alignment (CORAL) [31], and moment distance [22]. In addition, the metric to measure the similarity between domains can be implemented either explicitly as formulas as mentioned above, or implicitly as a generative adversarial neural network [29, 40, 44]. The latter form induces additional parameters for training and is challenging to harness label information to perform joint alignment between feature and label distribution [2], thus limiting its usage for large-scale and complex data. Besides the domain discrepancy between the target and each source, Ben-David et al. [10] derived a tighter bound for MSDA, where the relationship between pairwise sources should also be optimized. Thus, a couple of researches [18, 26] approximate this more compacted bound based on various metrics and implementations.

Despite the considerable number of approaches proposed to solve the MSDA problem under covariate shift assumption, authors [15, 27] point out that domain shift can also arise from label distribution mismatch, a setting that has been seldomly studied yet widely seen among data generation. To solve the existing challenge, both the practitioners [19, 24] assumed that the label shift was fully or partially available as a priori knowledge, while other authors [15, 27] avoid making such assumptions through computationally expensive estimation of label distribution.

2.2 Optimal transport

OT theory was initially introduced by mathematician Monge for investigating transporting resources from one site to another with minimal energy consumption. Nevertheless, solving the original OT problem is NP-hard as mentioned in [23]. Due to the relaxed and regularized formulations given by [1] and [6], respectively, this problem becomes solvable. Besides, a considerable number of real-world applications start to embed OT theory into their solutions, such as image processing [9], fluid mechanics [12], applied economics [23] and domain adaptation [6, 27].

OT theory mainly has twofold contributions to DA: the definition of transport plan that maps source data distribution to target space, and the derivation of Wasserstein distance for metricizing the similarity of two probability distributions even if their distributional supports are hardly overlapped [5]. To the best of our knowledge, only three OT-based methods exist under the context of unsupervised MSDA. Turrisi et al. [33] and Redko et al. [27] align the joint distribution $P(X, Y)$ of feature and label data based on OT distance. Since target labels are unavailable for training, the pseudo-labels on target domain are either generated by the progressively trained classifier or propagated from source domains. More recently, Montesuma et al. [21] estimate the target data distribution with a penalty originally proposed for semi-supervised OT. However, their approaches ignore the underlying impact of unbalanced class distribution and data correlation, thus hindering the models from being more robust and generative.

Our method to solve the unsupervised MSDA problem under class-wise imbalance and data correlation also depends on the OT theory. Compared with former approaches, we propose a designed metric CRWD in order to better utilize features and raw labels to form a rebalanced joint space, where the discrepancies between different domains can be metricized more accurately. Moreover, the devised metric is based on the enhanced Mahalanobis distance for multivariate decomposition and normalization. In addition, rather than introducing extra structure, *e.g.* a classifier or SVM, to produce pseudo-labels for benefiting from joint alignment, we rebalance the class proportion by simply modifying the optimal transport plan γ. The proposed method is presented in the following section.

3 Methodology

In this section, we will first give a formal definition of the MSDA problem. Then, we introduce the OT theory and its application in the context of DA briefly. In the end, this study describes the proposed CRWD in detail. The main notations used in this study are described in Table 1.

3.1 Problem definition

Suppose that we are given K source domains in total, where each source $D_{S_k} \in \{D_{S_1}, \dots, D_{S_K}\}$ consists of two components: samples $\mathbf{x}_{S_k}$ that are embedded in feature space $\mathcal{X}_{S_k}$, and their marginal probability distribution $\mu_{S_k} = P(\mathbf{x}_{S_k})$, *i.e.* $D_{S_k} = \{(\mathbf{x}_{S_k}, \mu_{S_k}) \mid k = 1, \dots, K\}$. Then, the corresponding task is defined in label space $\mathcal{Y}_{S_k} = \{1, \dots, i\}$ as a conditional probability distribution $P(\mathbf{y}_{S_k} \mid \mathbf{x}_{S_k})$ where $i \in \{1, \dots, M\}$ refers to the i^{th} class and $\mathbf{y}_{S_k} \in \mathcal{Y}_{S_k}$. Obviously, $P(\mathbf{y}_{S_k} \mid \mathbf{x}_{S_k})$ is usually represented as classifiers in machine learning. Similarly, we denote the target domain and task by $D_T = \{\mathbf{x}_T, \mu_T\}$ and $P(\mathbf{y}_T \mid \mathbf{x}_T)$, respectively. Under the setting of closed-set unsupervised MSDA, both feature and label spaces remain the same across multiple domains, namely, $\mathcal{X}_{S_1} = \dots = \mathcal{X}_{S_K} = \mathcal{X}_T$ and $\mathcal{Y}_{S_1} = \dots = \mathcal{Y}_{S_K} = \mathcal{Y}_T$. Besides, the true target labels $\mathbf{y}_T$ are unavailable during the training stage. However, for the standard setting of domain shift, the marginal distributions, together with the conditional ones, are inconsistent domain-wisely, that is, $\mu_{S_1} \neq \dots \neq \mu_{S_k} \neq \mu_{S_T}$ and $P(\mathbf{x}_{S_1} \mid \mathbf{y}_{S_1}) \neq \dots \neq P(\mathbf{x}_{S_k} \mid \mathbf{y}_{S_k}) \neq P(\mathbf{x}_T \mid \mathbf{y}_T)$. Furthermore, we suppose that label shift, i.e. $P(\mathbf{y}_{S_1}) \neq \dots \neq P(\mathbf{y}_{S_k}) \neq P(\mathbf{y}_T)$, is caused by the mismatching of class distribution, which can further represent domain data as:

$$\mu_S = \sum_{i=1}^{M} \phi_i \mu_i, \quad \mu_T = \sum_{i=1}^{M} \psi_i \mu_i, \tag{1}$$

Q2

Table 1 Notations and descriptions

Notation	Description	Notation	Description
K	Number of source domains	M	Number of classes
D_{S_k}, D_T	k^{th} source & target domain	X_D	Data from domain D
$\mathcal{X}_D$	Feature space on domain D	$\mathcal{Y}_D$	Label space on domain D
$\mathbf{x}_D$	Samples from domain D	$\mathbf{y}_D$	Labels from domain D
μ	Marginal distribution	$P(\cdot \mid \cdot)$	Conditional distribution
μ_i	Marginal distribution of class i	G, F	Generator & Classifier
ϵ	Statistical expectation	λ	Combined error
W	Wasserstein distance	γ	Optimal transport plan
C	Ground cost matrix	d_M	Mahalanobis distance
Σ	Covariance matrix	R	Rebalancing matrix
f_D	Class proportion of domain D	n	Number of samples
$< \cdot, \cdot >_F$	Frobenius dot product	$\lVert \cdot \rVert_p$	ℓ_p-norm

where μ_i is the marginal distribution given the i^{th} class, and the positive coefficients $\{\phi_i, \psi_i \mid \sum_{i=1}^{M}\phi_i = 1, \sum_{i=1}^{M}\psi_i = 1\}$ reweigh each class.

According to the analysis [28], given a loss function ℓ on domain D, the statistical expectation according to the samples $\mathbf{x}$ embedded in the distribution μ that a true hypothesis h diverges from learned model G_D and F_D is defined as:

$$\epsilon_D\left(h; G_D, F_D\right) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}\left[\ell\left(h(\mathbf{x}), F_D(G_D(\mathbf{x}))\right)\right]. \qquad (2)$$

From now on, we denote the source and target risks as shorthand $\epsilon_S(h; G_S, F_S) = \epsilon_S(h)$ and $\epsilon_T(h; G_T, F_T) = \epsilon_T(h)$, respectively. Based on [32], the goal of MSDA is to improve the performance of model F_D, G_D on target data X_T, with the knowledge extracted from annotated multiple source domains D_S and unlabeled target sample $\mathbf{x}_T$. Alternatively, the above-stated alignment process can be expressed as an optimization of the following generalization bound:

$$\epsilon_T\left(\hat{h}_S\right) \leq \epsilon_T\left(h_T^*\right) + 2\sum_{i=1}^{K}\alpha_i\left(d\left(\mu_{S_i}, \mu_T\right) + \lambda\right) + \mathcal{C}. \qquad (3)$$

In formula (3), $\hat{h}_S$ is the hypothesis estimated from multi-sources, $\epsilon_T\left(\hat{h}_S\right)$ is the error of implementing $\hat{h}_S$ upon the target domain; $\epsilon_T\left(h_T^*\right)$ is the loss of evaluating ground truth hypothesis h_T^* in the target domain; d is a certain divergence that measures the discrepancy between different probability distributions; λ is the combined error of the ideal model $h_{T,S}^*$ that minimizes both ϵ_S and ϵ_T; $\mathcal{C}$ denotes a constant, and α_i represent positive coefficients that sum to 1.

From this perspective, it can be concluded that the discrepancy d is closely related to class weights from (1). Moreover, we can find that minimizing both the distances between sources and target, namely, the term d and the combined error λ, can equally contribute to adaptation. In addition, this insight provides theoretical feasibility for our proposed work.

3.2 Optimal transport for domain adaptation

OT was essentially initiated to study the engineering problem of transporting piles of earth from one place to another with an optimal approach, *i.e.*, by minimizing the overall cost based on some predefined rules. As a result, the least effort of the whole transportation process is defined as Earth Mover's Distance(EMD), or Wasserstein distance.

In the context of DA, Wasserstein distance is introduced to search for an unknown transformation $T^* : \mathcal{X}_S \to \mathcal{X}_T$ that maps source data space to target one with the least energy cost. Generally, given the law of mass conservation $T_{\#}\mu_{D_1} = \mu_{D_2}$, the optimal transport plan T^* is based on the following problem:

$$T^* = \mathop{\arg\min}_{T}\sum C(\mathbf{x}_{D_1}, T(\mathbf{x}_{D_2})), \quad \text{s.t. } T_{\#}\mu_{D_1} = \mu_{D_2}, \qquad (4)$$

where C indicates a ground cost function defining the pairwise work of moving a probability mass from samples of one domain $\mathbf{x}_{D_1}$ to another $\mathbf{x}_{D_2} = T(\mathbf{x}_{D_1})$, and $T_{\#}$ denotes the push-forward operator. Kantorovich reformulated the original OT problem (4) as a linear program by relaxing the searching space into a collection of joint probability distribution Π with marginals μ_{D_1} and μ_{D_2}:

$$\begin{aligned}\gamma^* &= \mathop{\arg\min}_{\gamma\in\Pi(\mu_{\mathbf{D_1}}, \mu_{\mathbf{D_2}})}\sum_{i=1}^{n_{D_1}}\sum_{j=1}^{n_{D_2}}(C_{i,j}(\mathbf{x_{D_1}}, \mathbf{x_{D_2}})\cdot\gamma_{i,j})\\ &= \mathop{\arg\min}_{\gamma\in\Pi(\mu_{\mathbf{D_1}}, \mu_{\mathbf{D_2}})} < C, \gamma >_F, \quad s.t. \Pi\mathbb{1}_{n_{D_1}} = \mu_{D_1}, \Pi^{\top}\mathbb{1}_{n_{D_2}} = \mu_{D_2},\end{aligned} \qquad (5)$$

Q3

where $C_{i,j} = d_{i,j}(\mathbf{x_{D_1}}, \mathbf{x_{D_2}})$ is the element of ground cost matrix between pair of samples $\mathbf{x_{D_1}} \sim \mu_{D_1}$ and $\mathbf{x_{D_2}} \sim \mu_{D_2}$, and $< \cdot, \cdot >_F$ is the Frobenius dot product between cost matrix C and transport coupling γ with n_{D_1} and n_{D_2} being the number of samples.

Equation (5) has the dimension that scales quadratically with the size of input samples. As a result, the relaxed form is still infeasible for medium and large-scale applications. To overcome the existing drawback, Cuturi et al. added an entropy regularization term $H = -\gamma\log(\gamma)$ that forces the elements in matrix γ to distribute in a smoother way [1], allowing much faster Sinkhorn-Knopp approximation to be implemented. Additionally, Courty et al. [7] considered joint distributions of feature and label space for searching the optimal transport plan. However, since no labels from the target domain were available during the training stage, they proposed to apply a proxy strategy where labels are generated by the progressively trained model, *i.e.* $\hat{\mathbf{y}}_T = F(G(\mathbf{x}_T))$.

With the theory presented above, we write the general OT for SSDA framework as the following minimization problem:

$$\gamma^* = \mathop{\arg\min}_{\gamma} < C, \gamma >_F + \varepsilon\cdot H(\gamma), \qquad (6)$$

where the pairwise cost between feature and label distribution is presented by:

$$C = \beta\cdot d(G(\mathbf{x}_S), G(\mathbf{x}_T)) + \|\mathbf{y}_S, F(G(\mathbf{x}_T))\|_2, \qquad (7)$$

where ε and β are hyperparameters and d is the loss function measuring discrepancy between feature distributions. Conventionally, d is set to Euclidean distance. Once the optimal transport plan γ^* has been calculated according to (6), the Wasserstein distance can be simply computed as:

$$W(D_S, D_T) =< \gamma^*, C >_F . \tag{8}$$

Finally, the transported source distribution is provided by $\hat{X}_T = \gamma^* X_S$, which is known as Wasserstein barycentric projection.

3.3 Class-rebalanced Wasserstein distance

We propose two limitations of the framework (6). Initially, it combines the label information into the joint searching space via the label discrepancy in (7). However, the unbalanced class distribution may influence the geometry of underlying space, causing overfitting on the majorities and underfitting on the minorities of classes, which can inevitably lead to poor generalization and a class-biased network. The reason behind is that, with pure proxy strategy and data-sampling during the prevalent batch-training in machine learning, the pseudo-labels in the target domain are generated by the class-biased classifier in line with a limited mini-batch of source data X_T. Since the classes covered by this subset of data are usually fewer than the whole domain, the above strategy may yield unbalanced data and worsen the adaptation in a self-misleading way.

Secondly, while most OTDA methods employ Euclidean distance to calculate the ground cost, *i.e.* the loss d between extracted features, the performance of this metric degrades when metricizing correlative distributions from different domains and categories. Actually, correlated features may describe either similar or distinct objects, *e.g.*, photos of backpacks with or in car pattern and photos depicting real cars, as illustrated in Fig. 3. Based on another perspective [35], such data correlation suggests that samples belongs to one category may contain features from another, which is known as *conditional shift*. According to [8], such cases are extremely common in MSDA due to the size and comprehensiveness of datasets. Eventually, the two challenges mentioned above can generate negative transfer [48], upon which we elaborate our method for solutions as follows.

3.3.1 Class rectification

To alleviate the impact of the unbalanced class distribution, it is effective to utilize the class proportion of one domain *w.r.t.* and another as the coefficients to reweigh the transport plan γ, and thus the transport between the minority of classes is encouraged. Inspired by [27], we propose a strategy that first counts the samples of its own class to calculate the softmax-normalized proportion vector f_D. Before normalization, each sample $\mathbf{x}_i$ in domain D accounts for:

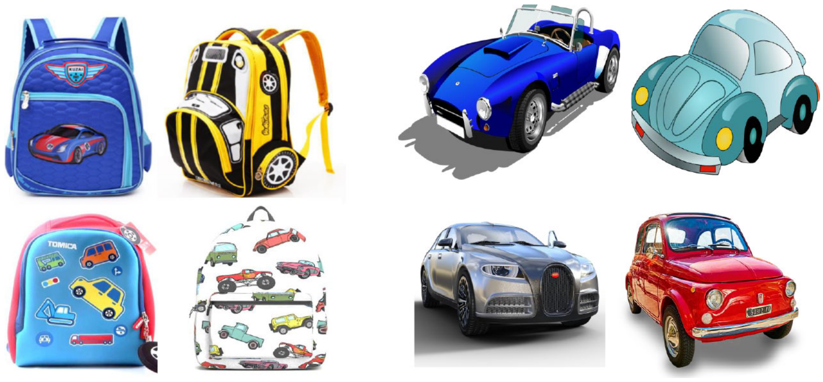
$$f_{D,i} = \frac{1}{\sum_{j=1}^{n_D} \mathbb{I}(\mathbf{y}_i = \mathbf{y}_j)}, \quad i = 1, \ldots, n_D \tag{9}$$

where n_D represents the number of samples, and $\mathbb{I}$ is the indicator function. For example, given $\mathbf{y} = (1, 1, 1, 2, 2, 3)$, the corresponding proportion is $softmax(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{2}, \frac{1}{2}, 1)$. Then, we compute the rebalance matrix R that combines the normalized proportions from two different domains D_i and D_j:

$$R = f_{D_i} \cdot f_{D_j}^{\top} . \tag{10}$$

Since labels from target domain are not available, we cannot obtain f_{D_j} by (9) if given $D_j = D_T$. Therefore, we assume that both samples and labels are projected into a latent space of joint distribution, from which the Wasserstein transport plan γ^* is optimized and the class structure is preserved. As a result, the pseudo-proportion $\hat{f}_{D_j}$ in case of unlabeled target domain D_j can be estimated by Wasserstein barycentric mapping, which is expressed as:

$$\hat{f}_{D_j} = \gamma^* f_{D_i} . \tag{11}$$



(a) Backpacks with/in car pattern (b) Cars

Fig. 3 Photos of the correlated features. It can be observed that both (backpack) and (car) contain car features. However, they are separately labeled as "Backpack" and "Car", indicating that features of backpacks and cars are correlated to a certain extent

By combining (10) and (11), the rectified transport plan is derived as follows for marginal distribution alignment:

$$\begin{aligned}\gamma^* &= \gamma^* \odot R \\ &= \gamma^* \odot (f_{D_i} \cdot f_{D_j}^\top) \\ &= \gamma^* \odot (f_{D_i} \cdot (\gamma^* f_{D_i})^\top) \\ &= \gamma^* \odot (f_{D_i} \cdot f_{D_i}^\top \cdot \gamma^{*\top}). \end{aligned} \tag{12}$$

where $\odot$ is the element-wise multiplication operator between matrices.

3.3.2 Principal component enhancement

To take the correlation between features into account, we consider the mixed-in features (e.g., car features in Fig. 3 backpack) as noisy principal component, as they can potentially confuse the classifier. And the polluted representative features (e.g., backpack features in Fig. 3 backpack) are treated as expected principal component, because these features are more consistent with labels. We enhance the principal components according to class distribution by exploiting the pair-wise cost d in (7). According to [23], arbitrary metric over a Riemann manifold could be implemented as this base metric. We consider constructing the ground cost on the more robust Mahalanobis distance rather than mere Euclidean distance. Based on the study [3], squared Mahalanobis distance deals with the challenge of correlation by principal component decomposition (or principal component analysis, PCA), which essentially contains a series of transformation including centralization, rotation and standardization. The squared Mahalanobis distance can be defined as:

$$d_M(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^\top A(\mathbf{x} - \mathbf{y}), \tag{13}$$

where $\mathbf{x}$ and $\mathbf{y}$ are observations, and A represents the inverse of the covariance matrix between $\mathbf{x}$ and $\mathbf{y}$, which must be positive semi-definite. In our approach, the observations are encoded feature vectors, *i.e.* $G(\mathbf{x}_S)$ and $G(\mathbf{x}_T)$. Specifically, the matrix A is designed as a Moor-Penrose inverse of the covariance matrix between extracted features for fast and stable computation, which is depicted as:

$$d_M(G(\mathbf{x}_S), G(\mathbf{x}_T)) = (G(\mathbf{x}_S) - G(\mathbf{x}_T))^\top A(G(\mathbf{x}_S) - G(\mathbf{x}_T)), \tag{14}$$

where $A = \Sigma^{-1}(G(\mathbf{x}_S), G(\mathbf{x}_T))$.

On top of the vanilla Mahalanobis distance for d as presented above, we also propose to introduce label information into this ground metric. The insight behind this is that for extracted pairs of highly correlated features, the principal components that can be exploited by Moor-Penrose inverse will be sparse, *i.e.*, inversing a nearly singular matrix Σ, causing more trivial zeros in A and a small base cost for d_M. According to [23], given a ground cost function d, the solution of OT exists if and only if the Lagrange dual f of the optimal plan is 1-Lipschitz continuous, that is:

$$\frac{\mid f(x) - f(y) \mid .}{d(x, y)} \le 1 \quad \text{or} \quad \mid f(x) - f(y) \mid \le d(x, y). \tag{15}$$

Therefore, in this case, the difference of dual f is restrained within a smaller range of $[0, d_M]$, indicating that the optimal transport plan γ tends to be over-flattened and less representative.

Hence, we assume that samples of minor classes are supposed to carry denser principal components, while the counterpart samples of major classes are more probably to be correlated. As a result, the ground cost is also rectified by a scaled rebalance matrix R' as follow:

$$d_M = R' \odot d_M, \tag{16}$$

where $R' = \tau R$, and τ is the scaling hyperparameter to ensure that the 1-Lipschitz condition is satisfied. Once the d_M between features of source and target data is calculated, the Wasserstein distance can then encourage the statistical transportation between the features of principal components rectified by the rebalance matrix R'. As a result, the expected components would be magnified and decoupled from the noisy ones. We display the workflow of principle component enhancement for backpack/car example in Fig. 4.

Compared with entropy regularization H, our scheme encourages transportations between features related to principal component, since the ultimate goal of CRWD is to extract informative features for adaptation. Obviously, methods such as [8] propose to metric-learn the Mahalanobis distance, which may violate the prerequisite of 1-Lipschitz continuity in the context of DA. Compared with their approximation, our approach is efficient, which still extracts sufficient and meaningful information concerning principal components for multi-domain transportation.

By combining (8), (12), (14), and (16), the overall domain distance between sources and target is written as:

$$W_{inter} = \sum_{i=1}^{K} W(D_{S_i}, D_T). \tag{17}$$

3.3.3 Conditional distribution alignment

According to the bound (3), only minimizing the domain discrepancy between sources and target may lead to insufficient adaptation. Therefore, we further measure and optimize the λ term in (3) as the intra-domain distances among multiple sources for better common feature extraction, that is:

$$W_{intra} = \sum_{i \neq j} W(D_{S_i}, D_{S_j}). \tag{18}$$

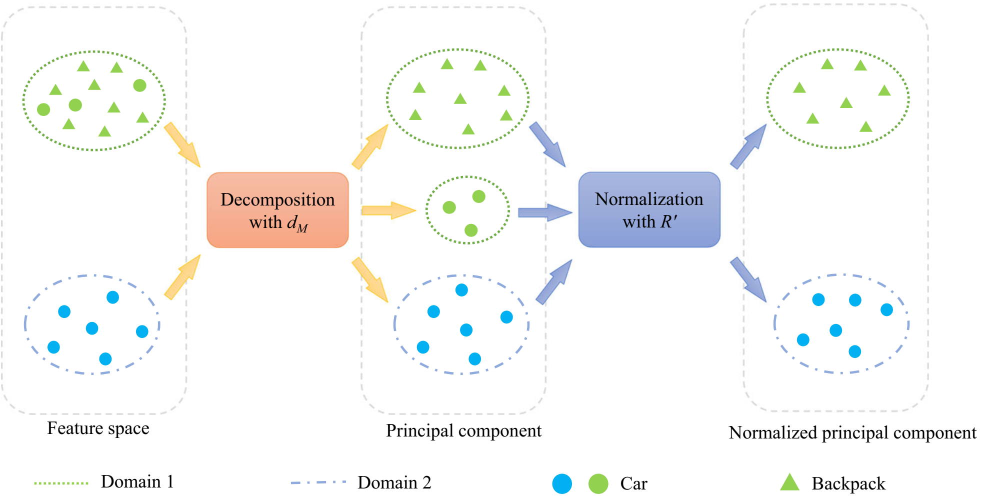


Fig. 4 The workflow of principle component enhancement

Obviously, we derived the rebalance matrix between annotated source and the unlabeled target domain in (10). Nevertheless, one can always perform rectification between two labeled sources by simply calculating:

$$\gamma^* = \gamma^* \odot R = \gamma^* \odot (f_{D_i} \cdot f_{D_j}^\top), \tag{19}$$

where f_{D_j} refers to the true class proportion of batched samples from another annotated source domain.

Finally, as suggested by the survey [32] to simultaneously align the marginal and conditional probability distribution, we follow the classifier alignment framework [30] in order to annotate the domain-invariant features extracted by the proposed distance. The alignment of conditional distribution includes the following three minimax steps:

(1) Simultaneously train a generator and two classifiers with CRWD to minimize classification loss.

$$\min_{G,F1,F2} L_{cls}(\mathbf{x}_S, \mathbf{y}_S) + W_{intra} + W_{inter}, \tag{20}$$

where L_{cls} is given by the cross-entropy between the predicted labels and the one-hot encoding of true labels on source domain.

(2) Train the pair of classifiers with fixed generator G to maximize the discrepancy between the classifiers, and thus the region of ambiguous features is specified as:

$$\min_{F1,F2} L_{cls}(\mathbf{x}_S, \mathbf{y}_S) - \|L_{F_1}(\mathbf{x}_T) - L_{F_2}(\mathbf{x}_T)\|_1, \tag{21}$$

where L_{F_1} and L_{F_2} are the output of the two classifiers.

(3) Re-train the generator G with fixed classifiers to let the model yield away from ambiguity detected in step 2.

$$\min_{G} \|L_{F_1}(\mathbf{x}_T) - L_{F_2}(\mathbf{x}_T)\|_1, \tag{22}$$

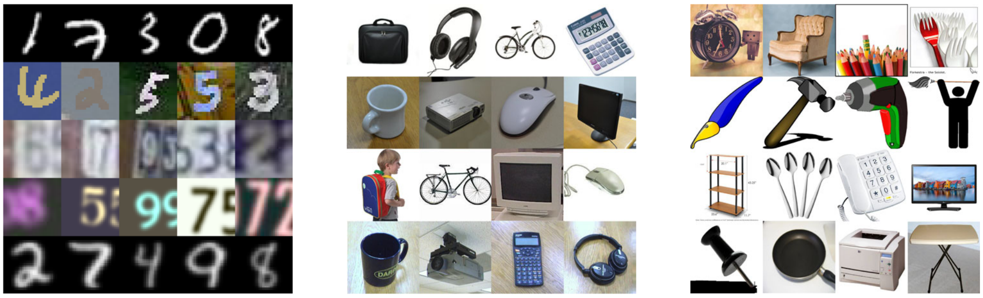
According to the original framework [30], after reaching max iteration, the two classifiers are similar enough to reach the optimality. We trivially pick the first trained classifier F_1 to predict target labels as:

$$\hat{\mathbf{y}}_T = F_1(G(\mathbf{x}_T)). \tag{23}$$

The pseudo-code of our algorithm is shown in the following Algorithm 1.

Algorithm 1 The CRWD algorithm.

```
Input: source samples $\mathbf{x}_S$, source labels $\mathbf{y}_S$, target samples $\mathbf{x}_T$
Output: target labels $\mathbf{y}_T$, trained generator $G$, trained classifier $F$
repeat
    for all $D_i \in D_S$, $D_j \in \{D_S, D_T\}$ and $D_i \neq D_j$ do
        Compute $f_{D_i}$ as (10);
        if $D_j \in D_S$ then
            Compute $f_{D_j}$ as (10);
        else if $D_j = D_T$ then
            Compute Euclidean-based transport plan $\gamma^*$ as (7)
            Compute $f_{D_j}$ as (12);
        end if
        Compute rebalancing matrix $R$ as (11);
        Rectify the ground cost $d$ by (20) with (18);
        Compute Mahalanobis-based transport plan $\gamma^*$ by (7);
        Rectify the transport plan $\gamma^*$ as (13) and (23);
        Calculate the domain distance $W$ as (9);
    end for
    Compute the classification loss as (24);
    Maximize the discrepancy between classifiers as (25);
    Minimize the loss between classifiers as (26);
until converge
Predict target label as (27);
```



(a) Digit-Five (b) Office-Caltech (c) Office-Home

Fig. 5 Photos sampled from 3 datasets. In (a), each row from top to bottom comes from MNIST, MNIST-M, SVHN, Synthetic Digits, and USPS, respectively. In (b), each row from top to bottom comes from Amazon, DSLR, Caltech, and Webcam, respectively. In (c), each row from top to bottom is from Art, Clipart, Product, and Real-World, separately

4 Experiments

This section first provides detailed information about the selected datasets together with the structure and implementation of our model. Then, we evaluate the proposed method via comparison with baselines and state-of-the-art MSDA methods on digit and image classification tasks. Finally, t-SNE embedding, ablation study and parameter sensitivity analysis are conducted to present the effectiveness of our model (Fig. 5).

4.1 Datasets

We select three standard MSDA datasets, *i.e.* Digit-five [22], Office-Caltech [42], and Office-Home [25], as our benchmarks. We give a detailed description as follows, and the summation is presented in Table 2. In addition, some samples are also demonstrated in Fig. 4.

(1) **Digit-five** contains five digit-related datasets sampled from different sources with ten categories, namely, *0˜9*. This benchmark includes handwritten MNIST, and USPS which are gray-processed, color-combined MNIST-M, street view house number SVHN, and computer-generated Synthetic Digits. By following the setting of [22], we sample 25,000 images for training and 9,000 for testing in MNIST, SVHN, MNIST-M, and Synthetic Digits as four domains, respectively. Since the USPS contains fewer digit data in comparison with all others, we take the full size of 9,298 images in it as the last domain.

(2) **Office-caltech** consists of the intersection of two different datasets office-31 and Caltech-256, where the samples belonging to shared categories are included. This benchmark has four domains that are crawled from the internet: Amazon contains images from shopping platform *amazon.com*, DSLR contains high-definition photos taken by digital cameras, Webcam contains low-quality photos taken by webcams, and Caltech includes pictures from comprehensive sources. The acquisition of images from each domain differs in various ways, such as angel, illumination condition, and device manufacturer. The dataset contains totally 2,533 images, where we take 70% of

Table 2 Dataset Information

Dataset	Number of class	Domain	Training/Testing split
Digit-five	10	MNIST(MT)	25,000/9,000
		MNIST-M(MM)	25,000/9,000
		SVHN(SV)	25,000/9,000
		Synthetic Digits(SD)	25,000/9,000
		USPS(US)	9,298/9,298
Office-Caltech	10	Amazon	671/287
		DSLR	110/47
		Caltech	207/88
		Webcam	910/389
Office-Home	65	Art	1,699/728
		Clipart	3,056/1,309
		Product	3,108/1,331
		Real-World(Real)	3,050/1,307

-The shorthands in parenthesis are used in Table 3, 5 them as the training set with the remaining part as the testing set.

(3) **Office-home** consists of 15,588 images with 65 classes from objects in office and home scenery. The whole dataset contains the following four different domains, respectively, Artistic images, Clip art, Product photos, and Real-World images. This dataset is particularly challenging for its large size and numerous categories. Similar to in Office-Caltech, we split the dataset by 70%/30% in the training and testing set.

To better visualize the existence of unbalanced class distribution, we calculate the number of samples belonging to the corresponding category for task →Amazon (see 4.2 for an explanation of this mark) in Fig. 6 as an example.

4.2 Implementation details

As illustrated in Fig. 2, CRWD is constructed with a feature extractor, alignment component, and two classifiers. The selection of networks is usually determined by transferring difficulty. Accordingly, we choose the popular Alexnet, ResNet-50, and ResNet-101 for Digit-five, Office-Caltech, and Office-Home as encoders, respectively. Besides, the ResNet-50 and ResNet-101 are pre-trained on ImageNet and then finetuned to approximate geometrical optimality. The architecture of our classifier is one fully-connected layer of artificial neurons, with the number of outputs being equal to classes per domain.

We choose one domain from the dataset as target during the training stage, and the others are left as multiple sources. Besides, we denote this form of transfer task as $\rightarrow D_T$, *e.g.*, →MNIST means to simultaneously transfer MNIST-M, SVHN, USPS, and Synthetic Digits to MNIST on Digit-five. On testing the model, we directly feed data from the target domain into the trained feature extractor and classifier. Next, we compute the cross-entropy loss between predicted labels and ground truth as accuracy. In order to demonstrate that the enhanced Mahalanobis distance is able to extract correlations among the multivariate of multiple domains, we train our model with pure Euclidean distance CRWD_{ℓ_2} and enhanced Mahalanobis distance CRWD_{maha} as the base cost, respectively.

Additionally, the parameters of β, ε are set to the recommended values from the original framework. Regarding the parameter τ, we search for the best value in the choice of $[10^{-2}, 10^{-1}, 1, 10, 10^{2}]$. The batch size is set to 128. All of our experiments are conducted on one Nvidia 2080ti GPU with PyTorch deep learning platform and POT toolbox [11]. We employ ADAM optimizer and an initial weight decay of 5×10^{-4}. To alleviate the impact of random batch-sampling, we test the trained model 5 times and calculate the mean value of these accuracies as the final result. To ensure convergence, the max epoch is set to 30.

By following [40], we use the source-combined baseline as a lower bound standard, where all sources are concatenated to form a single source domain. Then, we perform a traditional encoder-decoder paradigm between the combined source and target data. On evaluation, we compare CRWD with ABMSDA [49], MDAN [43], M3SDA [22], CMSS [41], MDDA [45], LtC-MSDA [34], MFSAN [47] , MADAN [46], MCD [30] and three OT-based method: WJDOT [33], JCPOT [27] and WBT [21].

4.3 Results

4.3.1 Accuracy evaluation

The comparison between CRWD and the selected baseline methods are reported in Tables 3, 4 and 5 on Digit-five, Office-Caltech, and Office-Home, respectively. Based on the averaged results (AR), we have the following conclusions:

(1) It can be seen from Tables 3, 4 and 5 that the average accuracies of the proposed method can outperform all other baseline methods on all datasets, with the digit recognition task achieving an accuracy of 94.1%, and object classification tasks on Office-Caltech and Office-Home datasets yield at 98.3% and 74.4%, which are higher than the best baseline methods by 2.3%, 1.9%, and 1.8%, respectively. Additionally, as shown in Table 4, the accuracy of our method exceeds

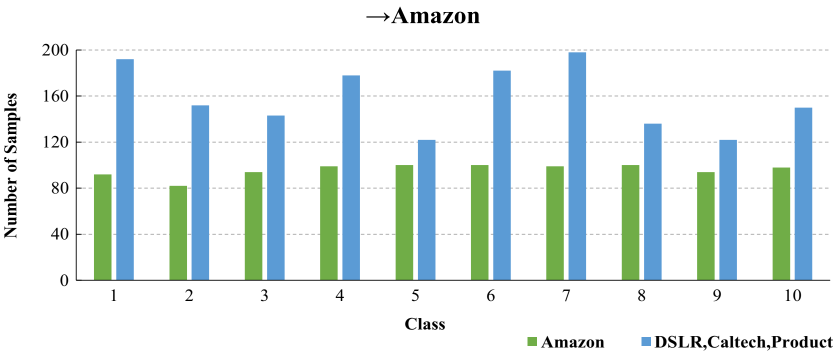


Fig. 6 The number of samples that belong to each class from source domains (in blue color) to target domain (in green color)

Table 3 Accuracy (%) on Digit-five

Method	→MM	→MT	→US	→SV	→SD	AR
source-combined	63.7	92.3	90.7	71.5	83.4	80.3
MDAN[43]	69.5	98.0	92.5	69.1	87.4	83.3
MCD[30]	72.5	96.2	95.3	78.9	87.5	86.1
M3SDA[22]	72.8	98.4	96.1	81.3	89.6	87.7
MDDA[45]	78.6	98.8	93.9	79.3	89.7	88.1
CMSS[41]	75.3	99.0	97.7	**88.4**	93.7	90.8
ABMSDA[49]	73.4	99.3	97.1	88.2	**97.7**	91.1
LtC-MSDA[34]	85.6	99.0	98.3	83.2	93.0	91.8
CRWD$_{\ell_2}$	89.2	99.5	**98.8**	86.5	93.4	93.5
CRWD$_{maha}$	**89.6**	**99.7**	**98.8**	87.8	94.7	**94.1**

three OT-based models JCPOT, WJDOT, and WBT by 14.3%, 5.9%, and 5.8%, respectively. The better performances carried by our strategy indicate that it acquires an advantage over different baselines.

(2) As discussed above, negative transfer may happen when a model fails to construct the correct relationship among multiple sources and target domains. Such occurrences are obvious on Office-Caltech and Office-Home, where the performances of JCPOT and MDAN are lower than source-combined bounds, respectively. Based on Table 3, we observe that CMSS can achieve the best result on →SVHN with an accuracy of 88.4%, while the performance drops to 75.3% on →MNIST-M. Similar phenomenon is WJDOT on →DSLR *w.r.t.* →Caltech. Comparatively, our method manages to rank top place on nearly all tasks. Therefore, we draw the conclusion that concerns need to be made upon label imbalance on MSDA, and that the effectiveness of the proposed work is able to mitigate this problem.

4.3.2 T-SNE embedding

To qualitatively illustrate the transferring ability of the proposed model before and after adaptation, in Fig. 7, we show the t-SNE embedding of CRWD compared with M3SDA on task →MNIST-M in Digit-five and on →Art in Office-Home.

As demonstrated in the first column of Fig. 7(a) and (b), the feature distribution before adaptation is quite chaotic and difficult for drawing boundaries to split different classes apart, visualizing the challenge of classification posed by MSDA. Based on the embeddings given by CRWD (third column) and M3SDA (second column), we make the following two observations: **i**) Both methods obtain clusters of different categories after adaptation, which justifies the necessity of utilizing cross-domain knowledge; **ii**) On both tasks, the clusters gathered by CRWD are more dense and compacted. However, they are relatively sparser in the second column of M3SDA. Typically, on task →MNIST-M for M3SDA, there remains an entangled cluster near the center, while the result for CRWD leaves no entanglement, suggesting that features provided by CRWD attain more domain-invariant knowledge than M3SDA. Henceforth, it can be concluded that the trained projection based on our method can effectively extract preferable discriminative features.

Table 4 Accuracy (%) on Office-Caltech

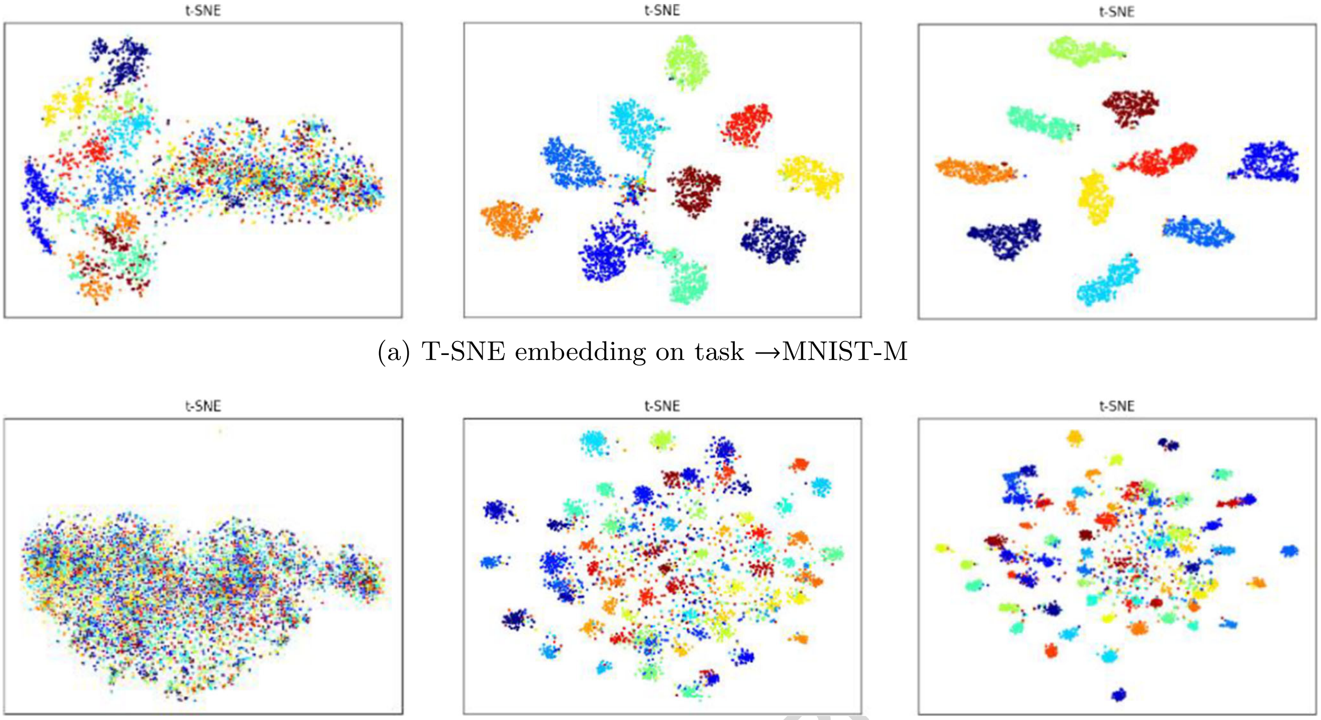
Method	→Amazon	→DSLR	→Webcam	→Caltech	AR
source-combined	90.6	96.8	88.4	83.0	89.7
JCPOT[27]	83.5	81.5	91.4	79.7	84.0
WJDOT[33]	94.2	**100.0**	89.3	85.9	92.4
WBT[21]	92.7	95.9	96.6	85.0	92.5
MDAN[43]	92.2	98.2	98.1	89.5	94.5
MCD[30]	92.1	99.1	99.5	91.5	95.6
M3SDA[22]	94.5	99.2	99.5	92.2	96.4
CRWD$_{\ell_2}$	97.2	99.6	99.8	95.5	98.0
CRWD$_{maha}$	**97.6**	99.8	**99.9**	**95.9**	**98.3**

4.3.3 Ablation study

We quantify the effectiveness of two major rebalancing components: the matrix R rectifying transport plan γ^*, and

Table 5 Accuracy (%) on Office-Home

Method	→Art	→Clipart	→Product	→Real	AR
source-combined	58.0	57.3	74.3	78.0	66.9
MDAN[43]	64.9	49.7	69.2	76.3	65.0
MADAN[46]	66.8	54.9	78.2	81.5	70.4
M3SDA[22]	64.1	**62.8**	76.2	78.6	70.4
MFSAN[47]	70.0	60.7	79.0	80.8	72.6
CRWD$_{\ell_2}$	71.1	57.8	81.7	83.1	73.4
CRWD$_{maha}$	**71.9**	59.2	**82.5**	**84.0**	**74.4**



(a) T-SNE embedding on task →MNIST-M

(b) T-SNE embedding on task →Art

Fig. 7 The t-SNE embedding of raw data (first column), M3SDA (second column) and CRWD (third column) on task →MNIST-M and →Art

the enhanced Mahalanobis distance *maha* serving as base metric in Wasserstein distance on benchmarks Digit-five and Office-Home.

According to Tables 6 and 7, we can observe that: **i**) Compared with utilizing neither of these components, implementing R or *maha* alone enhances the average accuracy by 2.1% and 1.4% on Digit-five, and by 2.3% and 1.0% on Office-Home, while matching these two components further boosts the performance by 2.7% and 3.3% on the two benchmarks, respectively; **ii**) Interestingly, on task →SVHN and →Synthetic Digits, the advantages of solely adding *maha*, i.e. 3.7% and 1.8%, are larger than only adding R, i.e. 3.4% and 1.0%, respectively. Considering the pictures from SVHN and Synthetic Digits often contains over one digit as shown in the third and fourth row in Fig. 5, the results imply that features on these two datasets are more correlated to each other, and the proposed scheme manages to extract such correlations.

4.3.4 Parameter sensitivity

As discussed in (16), the parameter τ in the supportive Mahalanobis distance is particularly important and needs specific adjustment. As a result, we fix other parameters and vary τ in the range of $[10^{-2}, 10^{-1}, 1, 10, 10^{2}]$ on Digit-five to reveal its influence. The result is shown in Fig. 8. Obviously, when τ is around 10, our model performs with the best accuracies in all tasks. The degradation before $\tau = 10$ can be explained by the undervalued ground metric, making CRWD less discriminative. In addition, when τ

Table 6 Ablation Study (%) on Digit-five

R	maha	→MM	→MT	→US	→SV	→SD	AR
		85.3	98.4	98.0	83.1	92.4	91.4
	✓	85.9	99.0	98.3	86.8	94.2	92.8
✓		89.2	99.5	**98.8**	86.5	93.4	93.5
✓	✓	**89.6**	**99.7**	**98.8**	**87.8**	**94.7**	**94.1**

Table 7 Ablation Study (%) on Office-Home

R	maha	→Art	→Clipart	→Product	→Real	AR
		68.2	55.2	79.2	81.6	71.1
	✓	69.0	56.4	80.3	82.7	72.1
✓		71.1	57.8	81.7	83.1	73.4
✓	✓	**71.9**	**59.2**	**82.5**	**84.0**	**74.4**

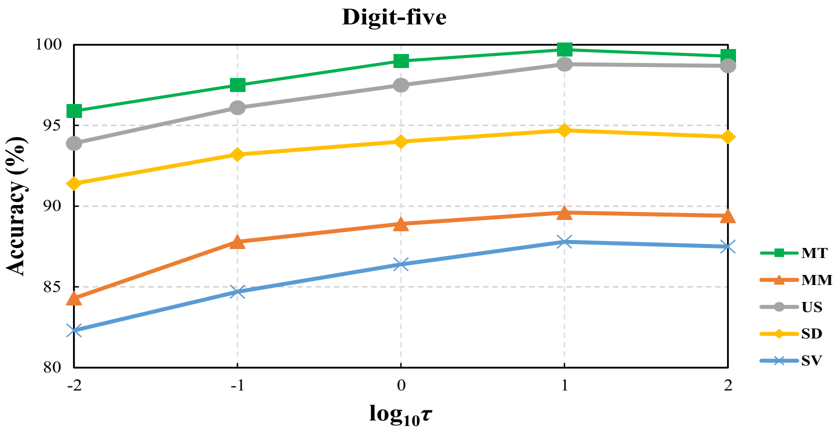


Fig. 8 The accuracy of CRWD with different τ on task →MNIST

is overvalued, the principal components are enhanced to saturation. Therefore, the accuracy is stabilized after $\tau = 10$. Similar trends can be found on other datasets.

5 Conclusion

To conclude, in this study, we present an end-to-end scheme, namely CRWD, for unsupervised MSDA under intrinsic biased label structures and data correlation. In order to better harness label information to construct representative joint distributions, our method bases on the OTDA framework, which is guided by class proportions produced from the latent Wasserstein space, by reweighing the transport plan and enhancing the base metric in the framework. The extensive experiments performed on Digit-five, Office-Caltech, and Office-Home benchmarks, which include real-world and synthetic datasets, demonstrate the effectiveness and robustness of the proposed method.

Currently, the Siamese feature extractor for CRWD is simply implemented as plain neural network, such as Alexnet or residual model. Nevertheless, the domain-invariant knowledge can be further exploited and extracted with additional attention mechanism, which has received considerable attention in recent years. Hopefully, by integrating the state-of-the-art attention technique, i.e., the Transformer, into CRWD, we can provide the classifier component with more representative common features across domains, and obtain a more effective model.

In addition, the results also demonstrate the potential of implementing our model on tasks with correlated and unbalanced data environments, typically, image processing for remote sensing, where images of different scenarios are imbalanced and multiple objects may exist in one picture. Since the manner of acquisition for remote sensing images varies from high resolution (RGB) to hyper-spectral (multi-channels), heterogeneous domain adaptation techniques need to be introduced into CRWD for handling multiple source data structured within different dimensions, remaining to be an interesting challenge.

Acknowledgements This work is supported by the Innovation Capacity Construction Project of Jilin Province Development and Reform Commission(2021FGWCXNLJSSZ10), the National Key Research and Development Program of China (No. 2020YFA0714103) and the Science & Technology Development Project of Jilin Province, China (20190302117GX), the Fundamental Research Funds for the Central Universities, JLU..

Data Availability The data that support the findings of this study are available from the corresponding author, upon reasonable request.

References

1. Altschuler J, Niles-Weed J, Rigollet P (2017) Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. Adv Neural Inf Process Syst, vol 30
2. Arora S, Ge R, Liang Y et al (2017) Generalization and equilibrium in generative adversarial nets (gans). In: International conference on machine learning. PMLR, pp 224–232
3. Brereton RG (2015) The mahalanobis distance and its relationship to principal component scores. J Chemom 29(3):143–145
4. Cao Y, Long M, Wang J (2018) Unsupervised domain adaptation with distribution matching machines. In: Proceedings of the AAAI conference on artificial intelligence
5. Chen Q, Liu Y, Wang Z et al (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
6. Courty N, Flamary R, Tuia D et al (2016) Optimal transport for domain adaptation. IEEE Trans Pattern Anal Mach Intell 39(9):1853–1865
7. Courty N, Flamary R, Habrard A et al (2017) Joint distribution optimal transportation for domain adaptation. Adv Neural Inf Process Syst, vol 30
8. Cuturi M, Avis D (2014) Ground metric learning. J Mach Learn Res 15(1):533–564
9. Damodaran BB, Flamary R, Seguy V et al (2020) An entropic optimal transport loss for learning deep neural networks under

label noise in remote sensing images. Comp Vision Image Underst 191:102,863

10. David SB, Lu T, Luu T et al (2010) Impossibility theorems for domain adaptation. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR workshop and conference proceedings, pp 129–136
11. Flamary R, Courty N, Gramfort A et al (2021) Pot: python optimal transport. J Mach Learn Res 22(78):1–8
12. Gangbo W, Li W, Osher S et al (2019) Unnormalized optimal transport. J Comput Phys 399:108,940
13. Gao P, Wu W, Li J (2021) Multi-source fast transfer learning algorithm based on support vector machine. Appl Intell:1–15
14. Guo J, Shah D, Barzilay R (2018) Multi-source domain adaptation with mixture of experts. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 4694–4703
15. Guo J, Gong M, Liu T et al (2020) Ltf: A label transformation framework for correcting label shift. In: International conference on machine learning, PMLR, pp 3843–3853
16. Hu C, Wang Y, Gu J (2020) Cross-domain intelligent fault classification of bearings based on tensor-aligned invariant subspace learning and two-dimensional convolutional neural networks. Knowl-Based Syst 209:106–214
17. Hu C, He S, Wang Y (2021) A classification method to detect faults in a rotating machinery based on kernelled support tensor machine and multilinear principal component analysis. Appl Intell 51(4):2609–2621
18. Li Y, Carlson DE et al (2018) Extracting relationships by multi-domain matching. Adv Neural Inf Process Syst, vol 31
19. Liu X, Guo Z, Li S et al (2021) Adversarial unsupervised domain adaptation with conditional and label shift: infer, align and iterate. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10,367–10,376
20. Mansour Y, Mohri M, Rostamizadeh A (2008) Domain adaptation with multiple sources. Adv Neural Inf Process Syst:21
21. Montesuma EF, Mboula FMN (2021) Wasserstein barycenter for multi-source domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16,785–16,793
22. Peng X, Bai Q, Xia X et al (2019) Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1406–1415
23. Peyré G, Cuturi M et al (2019) Computational optimal transport: with applications to data science. Foundations Trends Mach Learn 11(5-6):355–607
24. Podkopaev A, Ramdas A (2021) Distribution-free uncertainty quantification for classification under label shift. In: Uncertainty in artificial intelligence. PMLR, pp 844–853
25. Rahman MM, Fookes C, Baktashmotlagh M et al (2019) Multi-component image translation for deep domain generalization. In: IEEE winter conference on applications of computer vision (WACV). IEEE, pp 579-588
26. Rakshit S, Banerjee B, Roig G et al (2019) Unsupervised multi-source domain adaptation driven by deep adversarial ensemble learning. In: German conference on pattern recognition, Springer, pp 485–498
27. Redko I, Courty N, Flamary R et al (2019) Optimal transport for multi-source domain adaptation under target shift. In: The 22nd international conference on artificial intelligence and statistics. PMLR, pp 849–858
28. Redko I, Habrard A, Sebban M (2019) On the analysis of adaptability in multi-source domain adaptation. Mach Learn 108(8):1635–1652
29. Russo P, Tommasi T, Caputo B (2019) Towards multi-source adaptive semantic segmentation. In: International conference on image analysis and processing. Springer, pp 292–301
30. Saito K, Watanabe K, Ushiku Y et al (2018) Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3723–3732
31. Sun B, Feng J, Saenko K (2017) Correlation alignment for unsupervised domain adaptation. In: Domain adaptation in computer vision applications. Springer, pp 153–171
32. Sun S, Shi H, Wu Y (2015) A survey of multi-source domain adaptation. Inf Fusion 24:84–92
33. Turrisi R, Flamary R, Rakotomamonjy A et al (2020) Multi-source domain adaptation via weighted joint distributions optimal transport. arXiv:200612938
34. Wang H, Xu M, Ni B et al (2020) Learning to combine: knowledge aggregation for multi-source domain adaptation. In: European conference on computer vision. Springer, pp 727–744
35. Wang M, Deng W (2018) Deep visual domain adaptation: a survey. Neurocomputing 312:135–153
36. Wang Z, Jing B, Ni Y et al (2020) Adversarial domain adaptation being aware of class relationships. In: ECAI 2020. IOS Press, Santiago de Compostela, pp 1579-1586
37. Wilson G, Cook DJ (2020) A survey of unsupervised deep domain adaptation. ACM Trans Intell Syst Technol (TIST) 11(5): 1–46
38. Wu H, Yan Y, Ng MK et al (2020) Domain-attention conditional wasserstein distance for multi-source domain adaptation. ACM Trans Intell Syst Technol (TIST) 11(4):1–19
39. Wu H, Yan Y, Ye Y et al (2020) Geometric knowledge embedding for unsupervised domain adaptation. Knowl-Based Syst 191:105,155
40. Xu R, Chen Z, Zuo W et al (2018) Deep cocktail network: multi-source unsupervised domain adaptation with category shift. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3964–3973
41. Yang L, Balaji Y, Lim SN et al (2020) Curriculum manager for source selection in multi-source domain adaptation. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, 23–28 August, 2020, proceedings, Part XIV 16. Springer, pp 608–624
42. Zhang Y, Davison BD (2020) Impact of imagenet model selection on domain adaptation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops, pp 173–182
43. Zhao H, Zhang S, Wu G et al (2018) Adversarial multiple source domain adaptation. Adv Neural inf process syst 31:8559–8570
44. Zhao S, Li B, Yue X et al (2019) Multi-source domain adaptation for semantic segmentation. Adv Neural Inf Process Syst, vol 32
45. Zhao S, Wang G, Zhang S et al (2020) Multi-source distilling domain adaptation. In: Proceedings of the AAAI conference on artificial intelligence, pp 12,975–12,983
46. Zhao S, Li B, Xu P et al (2021) Madan: multi-source adversarial domain aggregation network for domain adaptation. Int J Comput Vis:1–26
47. Zhu Y, Zhuang F, Wang D (2019) Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In: Proceedings of the AAAI conference on artificial intelligence, pp 5989–5996
48. Zhuang F, Qi Z, Duan K et al (2020) A comprehensive survey on transfer learning. Proc IEEE 109(1):43–76
49. Zuo Y, Yao H, Xu C (2021) Attention-based multi-source domain adaptation. IEEE Trans Image Process 30:3793–3803

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES:

Q1. Author's Biography and photograph are desired. Please provide. Otherwise, please confirm if unnecessary.

Q2. Please check all equation if captured and presented correctly.

Q3. Please check equation 5 if correct.

Q4. Dummy citation for Figure 5 was inserted here. Please check if appropriate. Note that the order of main citations of figures in the text must be sequential.