Aggregating From Multiple Target-Shifted Sources

Changjian Shui[1] **Zijian Li**[2] **Jiaqi Li**[3] **Christian Gagné**[1 4] **Charles X. Ling**[3] **Boyu Wang**[3 5]

Abstract

Multi-source domain adaptation aims at leveraging the knowledge from multiple tasks for predicting a related target domain. A crucial aspect is to properly combine different sources based on their relations. In this paper, we analyzed the problem for aggregating source domains with different label distributions, where most recent source selection approaches fail. Our proposed algorithm differs from previous approaches in two key ways: the model aggregates multiple sources mainly through the similarity of semantic conditional distribution rather than marginal distribution; the model proposes a *unified* framework to select relevant sources for three popular scenarios, i.e., domain adaptation with limited label on target domain, unsupervised domain adaptation and label partial unsupervised domain adaption. We evaluate the proposed method through extensive experiments. The empirical results significantly outperform the baselines.

1. Introduction

Domain Adaptation (DA) (Pan & Yang, 2009) is based on the motivation that learning a new task is easier after having learned a similar task. By learning the inductive bias from a related source domain $\mathcal{S}$ and then leveraging the shared knowledge upon learning the target domain $\mathcal{T}$, the prediction performance can be significantly improved. Based on this, DA arises in tremendous deep learning applications such as computer vision (Zhang et al., 2019; Hoffman et al., 2018b), natural language processing (Ruder et al., 2019; Houlsby et al., 2019) and biomedical engineering (Raghu et al., 2019; Wang et al., 2020).

In various real-world applications, we want to transfer knowledge from *multiple sources* $(\mathcal{S}_1, \ldots, \mathcal{S}_T)$ to build a model for the target domain, which requires an effective selection and leveraging the *most useful* sources. Clearly, solely combining all the sources and applying one-to-one single DA algorithm can lead to undesired results, as it can include irrelevant or even untrusted data from certain sources, which can severely influence the performance (Zhao et al., 2020).

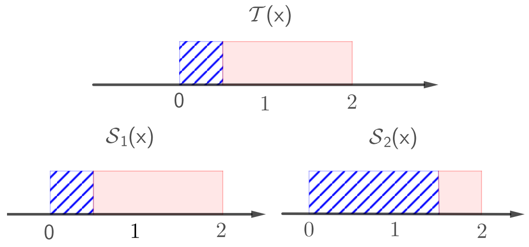


Figure 1. Limitation of merely considering marginal distribution $\mathbb{P}(x)$ in the source selection. In a binary classification, we have $\mathcal{S}_1(x) = \mathcal{S}_2(x) = \mathcal{T}(x)$, however adopting $\mathcal{S}_2$ is worse than $\mathcal{S}_1$ for predicting target $\mathcal{T}$ due to different decision boundaries.

To select related sources, most existing works (Zhao et al., 2018; Peng et al., 2019; Li et al., 2018a; Shui et al., 2019; Wang et al., 2019b; Wen et al., 2020) used the marginal distribution similarity $(\mathcal{S}_t(x), \mathcal{T}(x))$ to search the similar tasks. However, this can be problematic if their label distributions are different. As illustrated in Fig. 1, in a binary classification, the source-target marginal distributions are identical $(\mathcal{S}_1(x) = \mathcal{S}_2(x) = \mathcal{T}(x))$, however, using $\mathcal{S}_2$ for helping predict target domain $\mathcal{T}$ will lead to a negative transfer since their decision boundaries are rather different. This is not only theoretically interesting but also practically demanding. For example, in medical diagnostics, the disease distribution between the countries can be drastically different (Liu et al., 2004; Geiss et al., 2014). Thus applying existing approaches for leveraging related medical information from other data abundant countries to the destination country will be problematic.

In this work, we aim to address multi-source deep DA under different label distributions with $\mathcal{S}_t(y) \neq \mathcal{T}(y), \mathcal{S}_t(x|y) \neq \mathcal{T}(x|y)$, which is more realistic and challenging. In this case, if label information on $\mathcal{T}$ is absent (unsupervised DA), it is known as a underspecified problem and unsolvable *in*

[1]Université Laval [2]Guangdong University of Technology [3]Western University [4]Canada CIFAR AI Chair, Mila [5]Vector Institute. Correspondence to: Boyu Wang <bwang@csd.uwo.ca>, Christian Gagné <christian.gagne@gel.ulaval.ca>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

the general case (Ben-David et al., 2010; Johansson et al., 2019). For example, in Figure 1, it is impossible to know the preferable source if there is no label information on the target domain. Therefore, a natural extension is to assume limited label on target domain, which is commonly encountered in practice and a stimulating topic in recent research (Mohri & Medina, 2012; Wang et al., 2019a; Saito et al., 2019; Konstantinov & Lampert, 2019; Mansour et al., 2020). Based on this, we propose a novel DA theory with limited label on $\mathcal{T}$ (Theorem 1, 2), which motivates a novel source selection strategy by mainly considering the similarity of semantic conditional distribution $\mathbb{P}(x|y)$ and source re-weighted prediction loss.

Moreover, in the *specific case*, the proposed source aggregation strategy can be further extended to the unsupervised scenarios. Concretely, in our algorithm, we assume the problem satisfies the Generalized Label Shifted (GLS) condition (Combes et al., 2020), which is related to the cluster assumption and feasible in many practical applications, as shown in Sec. 5. Based on GLS, we simply add a label distribution ratio estimator, to assist the algorithm in selecting related sources in two popular multi-source scenarios: unsupervised DA and unsupervised label partial DA (Cao et al., 2018) with $\text{supp}(\mathcal{T}(y)) \subseteq \text{supp}(\mathcal{S}_t(y))$ (i.e., inherently label distribution shifted.)

Compared with previous work, the proposed method has the following benefits:

Better Source Aggregation Strategy We overcome the limitation of previous selection approaches when label distributions are different by significant improvements. Notably, the proposed approach is shown to simultaneously learn meaningful task relations and label distribution ratio.

Unified Method We provide a unified perspective to understand the source selection approach in different scenarios, in which previous approaches regarded them as separate problems. We show their relations in Fig. 2.

2. Related Work

Below we list the most related work and delegate additional related work in the Appendix.

Multi-Source DA has been investigated in previous literature with different aspects to aggregate source datasets. In the popular unsupervised DA, Zhao et al. (2018); Li et al. (2018b); Peng et al. (2019); Wen et al. (2020); Hoffman et al. (2018a) adopted the marginal distribution $d(\mathcal{S}_t(x), \mathcal{T}(x))$ of $\mathcal{H}$-divergence (Ben-David et al., 2007), discrepancy (Mansour et al., 2009) and Wasserstein distance (Arjovsky et al., 2017) to estimate domain relations. These works provided theoretical insights through upper bounding the target risk by the source risk, domain discrepancy of $\mathbb{P}(x)$ and an unobservable term η – the optimal risk on all the domains. However, as the counterexample indicates, relying on $\mathbb{P}(x)$ does not necessarily select the most related source. Therefore, Konstantinov & Lampert (2019); Wang et al. (2019a); Mansour et al. (2020) alternatively considered the divergence between two domains with limited target label by using $\mathcal{Y}$-discrepancy, which is commonly faced in practice and less focused in theory. However, we empirically show it is still difficult to handle target-shifted sources.

Target-Shifted DA (Zhang et al., 2013) is a common phenomenon in DA with $\mathcal{S}(y) \neq \mathcal{T}(y)$. Several theoretical analysis has been proposed under label shift assumption with $\mathcal{S}_t(x|y) = \mathcal{T}(x|y)$, e.g. Azizzadenesheli et al. (2019); Garg et al. (2020). Redko et al. (2019) proposed optimal transport strategy for the multiple unsupervised DA by assuming $\mathcal{S}_t(x|y) = \mathcal{T}(x|y)$. However, this assumption is restrictive for many real-world cases, e.g., in digits dataset, the conditional distribution is clearly different between MNIST and SVHN. In addition, the representation learning based approach is *not* considered in their framework. Therefore, Wu et al. (2019); Combes et al. (2020) analyzed DA under different assumptions in the *embedding space* $\mathcal{Z}$ for one-to-one unsupervised deep DA problem but did not provide guidelines of *leveraging different sources* to ensure a reliable transfer, which is our core contribution. Moreover, the aforementioned works focus on one specific scenario, without considering its flexibility for other scenarios such as *partial multi-source unsupervised DA*, where the label space in the target domain is a subset of the source domain (i.e., for some classes $\mathcal{S}_t(y) \neq 0$; $\mathcal{T}(y) = 0$) and class distributions are *inherently* shifted.

3. Problem Setup and Theoretical Insights

Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the output space. We consider the predictor h as a scoring function (Hoffman et al., 2018a) with $h : \mathcal{X} \times \mathcal{Y} \to R$ and predicted loss as $\ell : \mathbb{R} \to \mathbb{R}_+$ is positive, L-Lipschitz and upper bound by $L_{\max}$. We also assume that h is K-Lipschitz w.r.t. the feature x (given the same label), i.e. for $\forall y$, $\|h(x_1, y) - h(x_2, y)\|_2 \leq K\|x_1 - x_2\|_2$. We denote the expected risk w.r.t distribution $\mathcal{D}$: $R_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\,\ell(h(x, y))$ and its empirical counterpart (w.r.t. a given dataset $\hat{\mathcal{D}}$) $\hat{R}_{\mathcal{D}}(h) = \sum_{(x,y)\in\hat{\mathcal{D}}} \ell(h(x, y))$.

In this work, we adopt the commonly used Wasserstein distance as the metric to measure domains' similarity, which is theoretically tighter than the previously adopted TV distance (Gong et al., 2016) and Jensen-Shnannon divergence. Besides, based on previous work, a common strategy to adjust the imbalanced label portions is to introduce *label-distribution ratio* weighted loss with $R^{\alpha}_{\mathcal{S}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{S}}\,\alpha(y)\ell(h(x, y))$ with $\alpha(y) = \mathcal{T}(y)/\mathcal{S}(y)$. We also denote $\hat{\alpha}(y)$ as its empirical counterpart, estimated from the data.

Besides, in order to measure the task relations, we define $\boldsymbol{\lambda}$ ($\boldsymbol{\lambda}[t] \geq 0, \sum_{t=1}^{T} \boldsymbol{\lambda}[t] = 1$) as the *task relation coefficient* vector by assigning higher weight to the more related task. Then we prove Theorem 1, which proposes theoretical insights of combining source domains through properly estimating $\boldsymbol{\lambda}$.

Theorem 1 *Let* $\{\hat{\mathcal{S}}_t = \{(x_i, y_i)\}_{i=1}^{N_{\mathcal{S}_t}}\}_{t=1}^{T}$ *and* $\hat{\mathcal{T}} = \{(x_i, y_i)\}_{i=1}^{N_{\mathcal{T}}}$*, respectively be* T *source and target i.i.d. samples. For* $\forall h \in \mathcal{H}$ *with* $\mathcal{H}$ *the hypothesis family and* $\forall \boldsymbol{\lambda}$*, with high probability* $\geq 1 - 4\delta$*, the target risk can be upper bounded by:*

$$R_{\mathcal{T}}(h) \leq \underbrace{\sum_t \boldsymbol{\lambda}[t] \hat{R}^{\hat{\alpha}_t}_{\mathcal{S}_t}(h)}_{\text{(I)}} + \underbrace{L_{\max} d_{\infty}^{\sup} \sqrt{\sum_{t=1}^{T} \frac{\boldsymbol{\lambda}[t]^2}{\beta_t}} \sqrt{\frac{\log(1/\delta)}{2N}}}_{\text{(II)}}$$

$$+ \underbrace{LK \sum_t \boldsymbol{\lambda}[t] \mathbb{E}_{y \sim \hat{\mathcal{T}}(y)} W_1(\hat{\mathcal{T}}(x|Y=y) \| \hat{\mathcal{S}}_t(x|Y=y))}_{\text{(III)}}$$

$$+ \underbrace{L_{\max} \sup_t \|\alpha_t - \hat{\alpha}_t\|_2}_{\text{(IV)}} + \underbrace{\mathit{Comp}(N_{\mathcal{S}_1}, \ldots, N_{\mathcal{S}_T}, N_{\mathcal{T}}, \delta)}_{\text{(V)}},$$

where $N = \sum_{t=1}^{T} N_{\mathcal{S}_t}$ *and* $\beta_t = N_{\mathcal{S}_t}/N$ *and* $d_{\infty}^{\sup} = \max_{t \in [1,T], y \in \mathcal{Y}} \alpha_t(y)$ *the maximum true label distribution ratio value.* $W_1(\cdot \| \cdot)$ *is the Wasserstein-1 distance with* L_2*-distance as the cost function.* $\mathit{Comp}(N_{\mathcal{S}_1}, \ldots, N_{\mathcal{S}_T}, N_{\mathcal{T}}, \delta)$ *is a function that decreases with larger* $N_{\mathcal{S}_1}, \ldots, N_{\mathcal{T}}$*, given a fixed* δ *and hypothesis family* $\mathcal{H}$*. (See Appendix for details)*

Discussions (1) In (I) and (III), the relation coefficient $\boldsymbol{\lambda}$ is decided by $\hat{\alpha}_t$-weighted loss $\hat{R}^{\hat{\alpha}_t}_{\mathcal{S}_t}(h)$ and conditional Wasserstein distance $\mathbb{E}_{y \sim \hat{\mathcal{T}}(y)} W_1(\hat{\mathcal{T}}(x|Y=y) \| \hat{\mathcal{S}}_t(x|Y=y))$. Intuitively, a higher $\boldsymbol{\lambda}[t]$ is assigned to the source t with a *smaller weighted prediction loss* and *a smaller weighted semantic conditional Wasserstein distance*. In other words, the source selection depends on the similarity of the conditional distribution $\mathbb{P}(x|y)$ rather than $\mathbb{P}(x)$.

(2) If each source has equal samples ($\beta_t = 1/T$), then term (II) will become $\|\boldsymbol{\lambda}\|_2$, *a regularization term for the encouragement of uniformly leveraging all sources*. Term (II) is meaningful in the selection, because if several sources are simultaneously similar to the target, then the algorithm tends to select *a set of* related domains rather than only one most related domain (without regularization).

(3) Considering (I,II,III), we derive a novel source selection approach through the trade-off between assigning a higher $\boldsymbol{\lambda}[t]$ to the source t that has a smaller weighted prediction loss and similar semantic distribution with smaller conditional Wasserstein distance, and assigning balanced $\boldsymbol{\lambda}[t]$ for avoiding concentrating on one source.

(4) $\|\hat{\alpha}_t - \alpha_t\|_2$ (IV) indicates the gap between ground-truth and empirical label ratio. Therefore, if we can estimate a good label distribution ratio $\hat{\alpha}_t$, these terms can be small. Comp$(N_{\mathcal{S}_1}, \ldots, N_{\mathcal{S}_T}, N_{\mathcal{T}}, \delta)$ (V) is a function that reflects the convergence behavior, which decreases with larger observation numbers. If we fix $\mathcal{H}, \delta, N$ and $N_{\mathcal{T}}$, this term can be viewed as a constant.

Analysis in the Representation Learning Apart from Theorem 1, we further drive theoretical analysis in the *representation learning*, which motivates practical guidelines in the deep learning regime. We define a stochastic embedding g and we denote its conditional distribution w.r.t. latent variable Z (induced by g) as $\mathcal{S}(z|Y=y) = \int_x g(z|x)\mathcal{S}(x|Y=y)dx$. Then we have:

Theorem 2 *We assume the settings of loss, the hypothesis are the same with Theorem 1. We further denote the stochastic feature learning function* $g : \mathcal{X} \to \mathcal{Z}$*, and the hypothesis* $h : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$*. Then* $\forall \boldsymbol{\lambda}$*, the target risk is upper bounded by:*

$$R_{\mathcal{T}}(h, g) \leq \sum_t \boldsymbol{\lambda}[t] R^{\alpha_t}_{\mathcal{S}_t}(h, g)$$

$$+ LK \sum_t \boldsymbol{\lambda}[t] \mathbb{E}_{y \sim \mathcal{T}(y)} W_1(\mathcal{S}_t(z|Y=y) \| \mathcal{T}(z|Y=y)),$$

where $R_{\mathcal{T}}(h, g) = \mathbb{E}_{(x,y) \sim \mathcal{T}(x,y)} \mathbb{E}_{z \sim g(z|x)} \ell(h(z, y))$ *is the expected risk w.r.t. the function* g, h*.*

Theorem 2 motivates the practice of deep learning, which requires to learn an embedding function g that minimizes the weighted conditional Wasserstein distance and learn (g, h) that minimizes the weighted source risk $R^{\alpha_t}_{\mathcal{S}_t}$.

4. Practical Algorithm in Deep Learning

From the aforementioned theoretical results, we derive novel source aggregation approaches and training strategies, which can be summarized as follows.

Source Selection Rule Balance the trade-off between assigning a higher $\boldsymbol{\lambda}[t]$ to the source t that has a smaller weighted prediction loss and semantic conditional Wasserstein distance, and assigning balanced $\boldsymbol{\lambda}[t]$.

Training Rules (1) Learning an embedding function g that minimizes the weighted conditional Wasserstein distance, learning classifier h that minimizes the $\hat{\alpha}_t$-weighted source risk; (2) Properly estimate the label distribution ratio $\hat{\alpha}_t$.

Based on these ideas, we proposed Wasserstein Aggregation Domain Network (WADN) to automatically learn the network parameters and select related sources, where the high-level protocol is illustrated in Fig. 2.

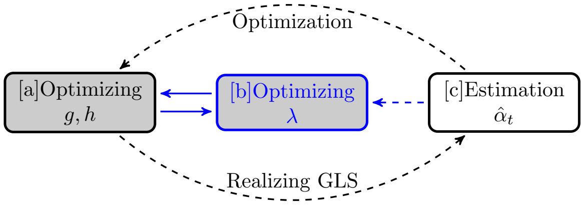


Figure 2. Illustration of proposed algorithm (WADN) and relation with other scenarios. WADN consists three components: **[a]** learning embedding function g and classifier h; **[b]** source aggregation through properly estimating $\boldsymbol{\lambda}$; **[c]** label distribution ratio ($\hat{\alpha}_t$) estimator. (1) If target labels are available, then WADN only requires **[a,b]** without gradually estimating $\hat{\alpha}_t$ (dashed arrows). (2) In the unsupervised scenarios, if we only have one source, WADN only contains **[a,c]** and recovers the single DA problem with label proportion shift, which can be solved under specific assumptions such as GLS (Li et al., 2019; Combes et al., 2020) or (Wu et al., 2019). (3) If there are multiple sources in the unsupervised DA, WADN gradually selects the related sources through interacting with other algorithmic components. (shown in blue).

4.1. Training Rules

Based on Theorem 2, given a fixed label ratio $\hat{\alpha}_t$ and fixed $\boldsymbol{\lambda}$, the goal is to find a representation function $g : \mathcal{X} \to \mathcal{Z}$ and a hypothesis function $h : \mathcal{Z} \times \mathcal{Y} \to \mathbb{R}$ such that:

$$\min_{g,h} \sum_t \boldsymbol{\lambda}[t] \hat{R}^{\hat{\alpha}_t}_{\mathcal{S}_t}(h, g) + C_0 \sum_t \boldsymbol{\lambda}[t] \mathbb{E}_{y\sim\hat{\mathcal{T}}(y)} W_1(\hat{\mathcal{S}}_t(z|Y = y) \| \hat{\mathcal{T}}(z|Y = y))$$

Explicit Conditional Loss One can *explicitly* solve the conditional optimal transport problem with g and h for a given $Y = y$. However, due to the high computational complexity in solving $T \times |\mathcal{Y}|$ optimal transport problems, the original form is practically intractable. To address this, we can approximate the conditional distribution on latent space Z as Gaussian distribution with identical Covariance matrix such that $\hat{\mathcal{S}}_t(z|Y = y) \approx \mathcal{N}(\mathbf{C}^y_t, \boldsymbol{\Sigma})$ and $\hat{\mathcal{T}}(z|Y = y) \approx \mathcal{N}(\mathbf{C}^y, \boldsymbol{\Sigma})$. Then we have $W_1(\hat{\mathcal{S}}_t(z|Y = y)\|\hat{\mathcal{T}}(z|Y = y)) \leq \|\mathbf{C}^y_t - \mathbf{C}^y\|_2$. Intuitively, the approximation term is equivalent to the well known *feature mean matching* (Sugiyama & Kawanabe, 2012), which computes the feature centroid of each class (on the latent space Z) and aligns them by minimizing their L_2 distance.

Implicit Conditional Loss Apart from approximation, we can derive a dual term for facilitating the computation, which is equivalent to the re-weighted Wasserstein adversarial loss by the label-distribution ratio.

Lemma 1 *The weighted conditional Wasserstein distance can be implicitly expressed as:*

$$\sum_t \boldsymbol{\lambda}[t] \mathbb{E}_{y\sim\mathcal{T}(y)} W_1(\mathcal{S}_t(z|Y = y) \| \mathcal{T}(z|Y = y))$$
$$= \max_{d_1,\cdots,d_T} \sum_t \boldsymbol{\lambda}[t] [\mathbb{E}_{z\sim\mathcal{S}_t(z)} \bar{\alpha}_t(z) d_t(z) - \mathbb{E}_{z\sim\mathcal{T}(z)} d_t(z)],$$

where $\bar{\alpha}_t(z) = \mathbf{1}_{\{(z,y)\sim\mathcal{S}_t\}} \alpha_t(Y = y)$, and $d_1, \ldots, d_T : \mathcal{Z} \to R_+$ are the 1-Lipschitz domain discriminators (Ganin et al., 2016).

Lemma 1 reveals that one can train T domain discriminators with weighted Wasserstein adversarial loss. When the source target distributions are identical, this loss recovers the conventional Wasserstein adversarial loss (Arjovsky et al., 2017). In practice, we adopt a hybrid approach by linearly combining the explicit and implicit matching, in which empirical results show its effectiveness.

Estimation $\hat{\alpha}$ When the target labels are available, $\hat{\alpha}_t$ can be directly estimated from the data with $\hat{\alpha}_t(y) = \hat{\mathcal{T}}(y)/\hat{\mathcal{S}}(y)$ and $\hat{\alpha}_t \to \alpha_t$ can be proved from asymptotic statistics. As for the unsupervised scenarios, we will discuss in Sec. 5.1.

4.2. Estimation Relation Coefficient $\boldsymbol{\lambda}$

Inspired by Theorem 1, given a *fixed* $\hat{\alpha}_t$ and (g, h), we estimate $\boldsymbol{\lambda}$ through optimizing the derived upper bound.

$$\min_{\boldsymbol{\lambda}} \quad \sum_t \boldsymbol{\lambda}[t] \hat{R}^{\hat{\alpha}_t}_{\mathcal{S}_t}(h, g) + C_1 \sqrt{\sum_{t=1}^{T} \frac{\boldsymbol{\lambda}^2[t]}{\beta_t}}$$
$$+ C_0 \sum_t \boldsymbol{\lambda}[t] \mathbb{E}_{y\sim\hat{\mathcal{T}}(y)} W_1(\hat{\mathcal{T}}(z|Y = y) \| \hat{\mathcal{S}}(z|Y = y))$$
$$\text{s.t} \quad \forall t, \boldsymbol{\lambda}[t] \geq 0, \sum_{t=1}^{T} \boldsymbol{\lambda}[t] = 1$$

In practice, $\hat{R}^{\hat{\alpha}_t}_{\mathcal{S}_t}(h, g)$ is the weighted empirical prediction loss and $\mathbb{E}_{y\sim\hat{\mathcal{T}}(y)} W_1(\hat{\mathcal{T}}(z|Y = y)\|\hat{\mathcal{S}}(z|Y = y))$ is approximated by the dynamic form of critic function from Lemma 1. Then, solving $\boldsymbol{\lambda}$ can be viewed as a standard convex optimization problem with linear constraints, which can be effectively resolved through standard convex optimizer.

5. Extension to Unsupervised Scenarios

In this section, we extend WADN to the unsupervised multi-source DA, which is known as unsolvable if semantic conditional distribution ($\mathcal{S}_t(x|y) \neq \mathcal{T}(x|y)$) and label distribution ($\mathcal{S}_t(y) \neq \mathcal{T}(y)$) are simultaneously different and no specific conditions are considered (Ben-David et al., 2010; Johansson et al., 2019).

In algorithm WADN, this challenging turns to properly estimate conditional Wasserstein distance and label distribution ratio $\hat{\alpha}_t(y)$ to help estimate $\boldsymbol{\lambda}$. According to Lemma 1, estimating the conditional Wasserstein distance can be viewed as $\hat{\alpha}_t$-weighted adversarial loss, thus if we can correctly estimate label distribution ratio such that $\hat{\alpha}_t \to \alpha_t$, then we can properly compute the conditional Wasserstein-distance through the adversarial term.

Therefore, the problem turns to properly estimate the label distribution ratio. To this end, we assume the problem satisfies Generalized Label Shift (GLS) condition (Combes et al., 2020), which has been theoretically justified and empirically evaluated in the single source unsupervised DA. The GLS condition states that *in unsupervised DA, there exists an optimal embedding function $g^\star \in \mathcal{G}$ that can ultimately achieve $\mathcal{S}_t(z|y) = \mathcal{T}(z|y)$ on the latent space.* (Combes et al., 2020) further pointed out that the clustering assumption (Chapelle & Zien, 2005) on $\mathcal{Z}$ is one sufficient condition to reach GLS, which is feasible for many practical applications.

Based on the achievability condition of GLS, the techniques of (Lipton et al., 2018; Garg et al., 2020) can be adopted to gradually estimate $\hat{\alpha}_t$ during learning the embedding function. Following this spirit, we add an distribution ratio estimator for $\{\hat{\alpha}_t\}_{t=1}^T$, shown in Sec. 5.1.

5.1. Estimation $\hat{\alpha}_t$

Unsupervised DA We denote $\bar{\mathcal{S}}_t(y)$, $\bar{\mathcal{T}}(y)$ as the predicted t-source/target label distribution through the hypothesis h, and also define $C_{\hat{\mathcal{S}}_t}[y,k] = \hat{\mathcal{S}}_t[\text{argmax}_{y'} h(z,y') = y, Y = k]$ is the t-source *prediction confusion matrix*. According to the GLS condition, we have $\bar{\mathcal{T}}(y) = \bar{\mathcal{T}}_{\hat{\alpha}_t}(y)$, with $\bar{\mathcal{T}}_{\hat{\alpha}_t}(Y=y) = \sum_{k=1}^{\mathcal{Y}} C_{\hat{\mathcal{S}}_t}[y,k]\hat{\alpha}_t(k)$ the constructed target prediction distribution from the t-source information. (See Appendix for justification). Then we can estimate $\hat{\alpha}_t$ through matching these two distributions by minimizing $D_{\text{KL}}(\bar{\mathcal{T}}(y)\|\bar{\mathcal{T}}_{\hat{\alpha}_t}(y))$, which is equivalent to solve the following convex optimization:

$$\begin{aligned} \min_{\hat{\alpha}_t} \quad & -\sum_{y=1}^{|\mathcal{Y}|} \bar{\mathcal{T}}(y) \log(\sum_{k=1}^{|\mathcal{Y}|} C_{\hat{\mathcal{S}}_t}[y,k]\hat{\alpha}_t(k)) \\ \text{s.t} \quad & \forall y \in \mathcal{Y}, \hat{\alpha}_t(y) \geq 0, \quad \sum_{y=1}^{|\mathcal{Y}|} \hat{\alpha}_t(y)\hat{\mathcal{S}}_t(y) = 1 \end{aligned} \tag{1}$$

Unsupervised Partial DA If we have $\text{supp}(\mathcal{T}(y)) \subseteq \text{supp}(\mathcal{S}_t(y))$, α_t will be sparse due to the non-overlapped classes. Thus, we impose such prior knowledge by adding a regularizer $\|\hat{\alpha}_t\|_1$ to the objective of Eq. (1) to induce the sparsity in $\hat{\alpha}_t$.

In training the neural network, the non-overlapped classes will be automatically assigned with a small or zero $\hat{\alpha}_t$, then (g,h) will be less affected by the classes with small $\hat{\alpha}_t$.

Algorithm 1 WADN (unsupervised scenario, one epoch)

```
Ensure: Label ratio $\hat{\alpha}_t$ and task relation $\boldsymbol{\lambda}$. Feature Learner
    $g$, Classifier $h$, statistic critic function $d_1, \dots, d_T$, class
    centroid for source $\mathbf{C}_t^y$ and target $\mathbf{C}^y$.$(t = 1, \dots, T)$
▷ DNN Parameter Training Stage (fixed $\alpha_t$ and $\boldsymbol{\lambda}$) ◁
for mini-batch of samples $(\mathbf{x}_{\mathcal{S}_1}, \mathbf{y}_{\mathcal{S}_1}) \sim \hat{\mathcal{S}}_1, \dots,$
    $(\mathbf{x}_{\mathcal{S}_T}, \mathbf{y}_{\mathcal{S}_T}) \sim \hat{\mathcal{S}}_T$, $(\mathbf{x}_{\mathcal{T}}) \sim \hat{\mathcal{T}}$ do
    Target predicted-label $\bar{\mathbf{y}}_{\mathcal{T}} = \text{argmax}_y h(g(\mathbf{x}_{\mathcal{T}}), y)$
    Compute unnormalized source confusion matrix on
        current batch $C_{\hat{\mathcal{S}}_t}[y,k]$.
    Compute feature centroid for source $C_t^y$ and target
        $C^y$ on current batch; Use moving average to update
        source and target class centroid $\mathbf{C}_t^y$ and $\mathbf{C}^y$.
    Updating $g, h, d_1, \dots, d_T$, by optimizing:

        $\min_{g,h} \max_{d_1,\dots,d_T} \underbrace{\sum_t \boldsymbol{\lambda}[t]\hat{R}^{\hat{\alpha}_t}_{\mathcal{S}_t}(h,g)}_{\text{Classification Loss}}$
        $+ \epsilon C_0 \underbrace{\sum_t \boldsymbol{\lambda}[t]\mathbb{E}_{y\sim\bar{\mathcal{T}}(y)}\|\mathbf{C}_t^y - \mathbf{C}^y\|_2}_{\text{Explicit Conditional Loss}}$
        $+(1-\epsilon)C_0 \underbrace{\sum_t \boldsymbol{\lambda}[t][\mathbb{E}_{z\sim\hat{\mathcal{S}}_t(z)}\bar{\alpha}_t(z)d(z) - \mathbb{E}_{z\sim\hat{\mathcal{T}}(z)}d(z)]}_{\text{Implicit Conditional Loss}}$

end for
▷ Estimation $\hat{\alpha}_t$ and $\boldsymbol{\lambda}$ ◁
Compute normalized source confusion matrix; Solve
    $\{\hat{\alpha}_t\}_{t=1}^T$ w.r.t. current training epoch through Sec.5.1 ;
    Update global $\hat{\alpha}_t$ through moving average.
Solve $\boldsymbol{\lambda}$ through Sec.4.2 w.r.t. current training epoch;
    Update global $\boldsymbol{\lambda}$ through moving average.
```

5.2. Algorithm implementation and discussion

We give an algorithmic description of Fig. 2, shown in Algorithm 1. The high-level protocol is to *iteratively* optimizes the neural-network parameters to gradually realize GLS condition with $g \to g^\star$ and dynamically update $\boldsymbol{\lambda}$, $\hat{\alpha}_t$ to better estimate conditional distance and aggregate the sources. The GLS assumes the achievability of existing an optimal $g^\star$. Our iterative algorithm can achieve a stationary solution but due to the highly non-convexity of deep network, converging to the global optimal does not necessarily guarantee.

Concretely, we update the $\hat{\alpha}_t$ and $\boldsymbol{\lambda}$ on the fly through a moving averaging strategy. Within one training epoch over the mini-batches, we fix the $\hat{\alpha}_t$ and $\boldsymbol{\lambda}$ and optimize the network parameters g, h. Then at each training epoch, we re-estimate the $\hat{\alpha}_t$ and $\boldsymbol{\lambda}$ by using the proposed estimator. When computing the explicit conditional loss, we empiri-

Table 1. Unsupervised DA: Accuracy (%) on *Source-Shifted* Amazon Review (Left) and Digits (Right).

Target	Books	DVD	Electronics	Kitchen	Average
Source	68.15$_{\pm 1.37}$	69.51$_{\pm 0.74}$	82.09$_{\pm 0.88}$	75.30$_{\pm 1.29}$	73.81
DANN	65.59$_{\pm 1.35}$	67.23$_{\pm 0.71}$	80.49$_{\pm 1.11}$	74.71$_{\pm 1.53}$	72.00
MDAN	68.77$_{\pm 2.31}$	67.81$_{\pm 2.46}$	80.96$_{\pm 0.77}$	75.67$_{\pm 1.96}$	73.30
MDMN	70.56$_{\pm 1.05}$	69.64$_{\pm 0.73}$	82.71$_{\pm 0.71}$	77.05$_{\pm 0.78}$	74.99
M^{3}SDA	69.09$_{\pm 1.26}$	68.67$_{\pm 1.37}$	81.34$_{\pm 0.66}$	76.10$_{\pm 1.47}$	73.79
DARN	71.21$_{\pm 1.16}$	68.68$_{\pm 1.12}$	81.51$_{\pm 0.81}$	77.71$_{\pm 1.09}$	74.78
WADN	**73.72**$_{\pm 0.63}$	**79.64**$_{\pm 0.34}$	**84.64**$_{\pm 0.48}$	**83.73**$_{\pm 0.50}$	**80.43**

Target	MNIST	SVHN	SYNTH	USPS	Average
Source	84.93$_{\pm 1.50}$	67.14$_{\pm 1.40}$	78.11$_{\pm 1.31}$	86.02$_{\pm 1.12}$	79.05
DANN	86.99$_{\pm 1.53}$	69.56$_{\pm 2.26}$	78.73$_{\pm 1.30}$	86.81$_{\pm 1.74}$	80.52
MDAN	87.86$_{\pm 2.24}$	69.13$_{\pm 1.56}$	79.77$_{\pm 1.69}$	86.50$_{\pm 1.59}$	80.81
MDMN	87.31$_{\pm 1.88}$	69.84$_{\pm 1.59}$	80.27$_{\pm 0.88}$	86.61$_{\pm 1.41}$	81.00
M^{3}SDA	87.22$_{\pm 1.70}$	68.89$_{\pm 1.93}$	80.01$_{\pm 1.77}$	86.39$_{\pm 1.68}$	80.87
DARN	86.98$_{\pm 1.29}$	68.59$_{\pm 1.79}$	80.68$_{\pm 0.61}$	86.85$_{\pm 1.78}$	80.78
WADN	**89.07**$_{\pm 0.72}$	**71.66**$_{\pm 0.77}$	**82.06**$_{\pm 0.89}$	**90.07**$_{\pm 1.10}$	**83.22**

Table 2. Unsupervised DA: Accuracy (%) on Office-Home

Target	Art	Clipart	Product	Real-World	Average
Source	49.25$_{\pm 0.60}$	46.89$_{\pm 0.61}$	66.54$_{\pm 1.72}$	73.64$_{\pm 0.91}$	59.08
DANN	50.32$_{\pm 0.32}$	50.11$_{\pm 1.16}$	68.18$_{\pm 1.27}$	73.71$_{\pm 1.63}$	60.58
MDAN	67.93$_{\pm 0.36}$	66.61$_{\pm 1.32}$	79.24$_{\pm 1.52}$	81.82$_{\pm 0.65}$	73.90
MDMN	68.38$_{\pm 0.58}$	67.42$_{\pm 0.53}$	82.49$_{\pm 0.56}$	83.32$_{\pm 1.93}$	75.28
M^{3}SDA	63.77$_{\pm 1.07}$	62.30$_{\pm 0.44}$	75.85$_{\pm 1.24}$	79.92$_{\pm 0.60}$	70.46
DARN	69.89$_{\pm 0.42}$	68.61$_{\pm 0.50}$	83.37$_{\pm 0.62}$	84.29$_{\pm 0.46}$	76.54
WADN	**73.78**$_{\pm 0.43}$	**70.18**$_{\pm 0.54}$	**86.32**$_{\pm 0.38}$	**87.28**$_{\pm 0.87}$	**79.39**

cally adopt the target pseudo-label. The implicit and explicit trade-off coefficient is set as $\epsilon = 0.5$. As for optimization $\boldsymbol{\lambda}$ and α_t, it is a standard convex optimization problem and we use package CVXPY.

As for WADN with limited target label, we do not require label distribution ratio component and directly compute $\hat{\alpha}_t$.

6. Experiments

In this section, we compare the proposed approaches with several baselines on the popular tasks. For all the scenarios, the following multi-source DA baselines are evaluated: (I) **Source** method applied only labelled source data to train the model. (II) **DANN** (Ganin et al., 2016). We follow the protocol of (Wen et al., 2020) to merge all the source dataset as a global source domain. (III) **MDAN** (Zhao et al., 2018); (IV) **MDMN** (Li et al., 2018b); (V) **M^{3}SDA** (Peng et al., 2019) adopted maximizing classifier discrepancy (Saito et al., 2018) and (VI) **DARN** (Wen et al., 2020). For the multi-source with limited target label and partial unsupervised multi-source DA, we additionally add specific baselines. All the baselines are re-implemented in the same network structure for fair comparisons. The detailed network structures, hyper-parameter settings, training details are delegated in Appendix.

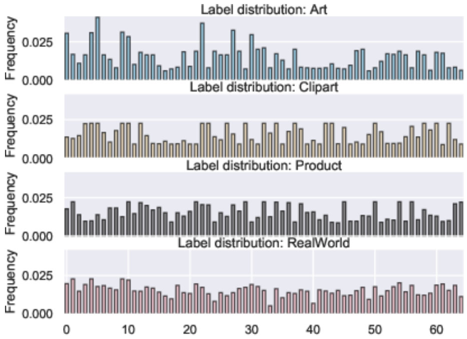


Figure 3. Label distribution on Office-Home Dataset

We evaluate the performance on three different datasets: (1) **Amazon Review.** (Blitzer et al., 2007) It contains four domains (Books, DVD, Electronics, and Kitchen) with positive and negative product reviews. We follow the common data pre-processing strategies as (Chen et al., 2012) to form a 5000-dimensional bag-of-words feature. Note that the label distribution in the original dataset is uniform. *To show the benefits of the proposed approach, we create a label distribution drifted task by randomly dropping* 50% *negative reviews of all the sources while keeping the target identical.* (2) **Digits**. It consists four digits recognition datasets including MNIST, USPS (Hull, 1994), SVHN (Netzer et al., 2011) and Synth (Ganin et al., 2016). *We also create a label distribution drift for the sources by randomly dropping* 50% *samples on digits 5-9 and keep target identical.* (3) **Office-Home Dataset** (Venkateswara et al., 2017). It contains 65 classes for four different domains: Art, Clipart, Product and Real-World. We used the ResNet50 (He et al., 2016) pretrained from the ImageNet in PyTorch as the base network for feature learning and put a MLP for the classification. The label distributions in these four domains are different and we did not manually create a label drift, shown in Fig. 3.

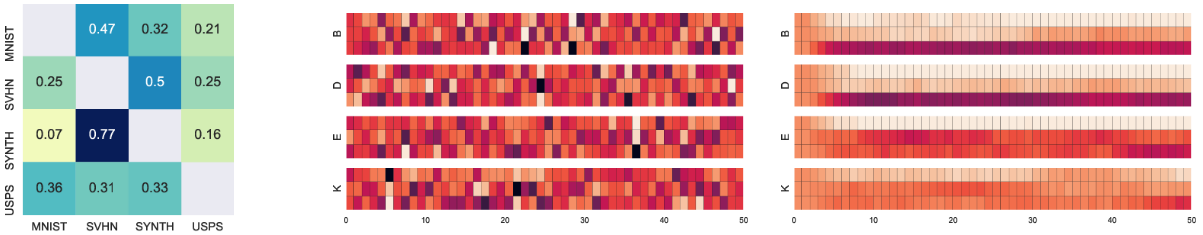
6.1. Unsupervised Multi-Source DA

In the unsupervised multi-source DA, we evaluate the proposed approach on all three datasets. We use a similar hyper-parameter selection strategy as in DANN (Ganin et al., 2016). All reported results are averaged from five runs. The detailed experimental settings are illustrated in Appendix.

Table 3. Multi-Source DA with Limited Target Label: Accuracy $(\%)$ on *Source-Shifted* Amazon Review (Left) and Digits (Right).

Target	Books	DVD	Electronics	Kitchen	Average
Source + Tar	72.59$_{\pm 1.89}$	73.02$_{\pm 1.84}$	81.59$_{\pm 1.58}$	77.03$_{\pm 1.73}$	76.06
DANN	67.35$_{\pm 2.28}$	66.33$_{\pm 2.42}$	78.03$_{\pm 1.72}$	74.31$_{\pm 1.71}$	71.50
MDAN	68.70$_{\pm 2.99}$	69.30$_{\pm 2.21}$	78.78$_{\pm 2.21}$	74.07$_{\pm 1.89}$	72.71
MDMN	69.19$_{\pm 2.09}$	68.71$_{\pm 2.39}$	81.88$_{\pm 1.46}$	78.51$_{\pm 1.91}$	74.57
M^{3}SDA	69.28$_{\pm 1.78}$	67.40$_{\pm 0.46}$	76.28$_{\pm 0.81}$	76.50$_{\pm 1.19}$	72.36
DARN	68.57$_{\pm 1.35}$	68.77$_{\pm 1.81}$	80.19$_{\pm 1.66}$	77.51$_{\pm 1.20}$	73.76
RLUS	71.83$_{\pm 1.71}$	69.64$_{\pm 2.39}$	81.98$_{\pm 1.04}$	78.69$_{\pm 1.15}$	75.54
MME	69.66$_{\pm 0.58}$	71.36$_{\pm 0.96}$	78.88$_{\pm 1.51}$	76.64$_{\pm 1.73}$	74.14
WADN	**74.83**$_{\pm 0.84}$	**75.05**$_{\pm 0.62}$	**84.23**$_{\pm 0.58}$	**81.53**$_{\pm 0.90}$	**78.91**

Target	MNIST	SVHN	SYNTH	USPS	Average
Source + Tar	79.63$_{\pm 1.74}$	56.48$_{\pm 1.90}$	69.64$_{\pm 1.38}$	86.29$_{\pm 1.56}$	73.01
DANN	86.77$_{\pm 1.30}$	69.13$_{\pm 1.09}$	78.82$_{\pm 1.35}$	86.54$_{\pm 1.03}$	80.32
MDAN	86.93$_{\pm 1.05}$	68.25$_{\pm 1.53}$	79.80$_{\pm 1.17}$	86.23$_{\pm 1.41}$	80.30
MDMN	77.59$_{\pm 1.36}$	69.62$_{\pm 1.26}$	78.93$_{\pm 1.64}$	87.26$_{\pm 1.13}$	78.35
M^{3}SDA	85.88$_{\pm 2.06}$	68.84$_{\pm 1.05}$	76.29$_{\pm 0.95}$	87.15$_{\pm 1.10}$	79.54
DARN	86.58$_{\pm 1.46}$	68.86$_{\pm 1.30}$	80.47$_{\pm 0.67}$	86.80$_{\pm 0.89}$	80.68
RLUS	87.61$_{\pm 1.08}$	**70.50**$_{\pm 0.94}$	79.52$_{\pm 1.30}$	86.70$_{\pm 1.13}$	81.08
MME	87.24$_{\pm 0.95}$	65.20$_{\pm 1.35}$	80.31$_{\pm 0.60}$	87.88$_{\pm 0.76}$	80.16
WADN	**88.32**$_{\pm 1.17}$	**70.64**$_{\pm 1.02}$	**81.53**$_{\pm 1.11}$	**90.53**$_{\pm 0.71}$	**82.75**



(a) Visualization of $\boldsymbol{\lambda}$

(b) DARN (Wen et al., 2020)

(c) WADN

Figure 4. Understanding Aggregation Principles in Unsupervised DA. (a) Visualization of $\boldsymbol{\lambda}$ on digits datset, each row corresponds to a target domain, which indicates a *non-uniform* and *non-symmetric* task relations. (b,c) The evolution of $\boldsymbol{\lambda}$ with three sources of Amazon dataset (B=Books, D=DVD, E=Electronics, K=Kitchen) during the training epoch. We compare with a recent principle approach DARN, which uses $\mathbb{P}(x)$ to measure the similarity and dynamically update the $\boldsymbol{\lambda}$. The results verifies the limitation of DARN under changing label distributions with relative unstable results.

The empirical results are illustrated in Tab. 1 and 2. Since we did not change the target label distribution throughout the whole experiment, we still report the target accuracy as the metric. We report the means and standard deviations for each approach. The best approaches based on a two-sided Wilcoxon signed-rank test (significance level $p = 0.05$) are shown in bold.

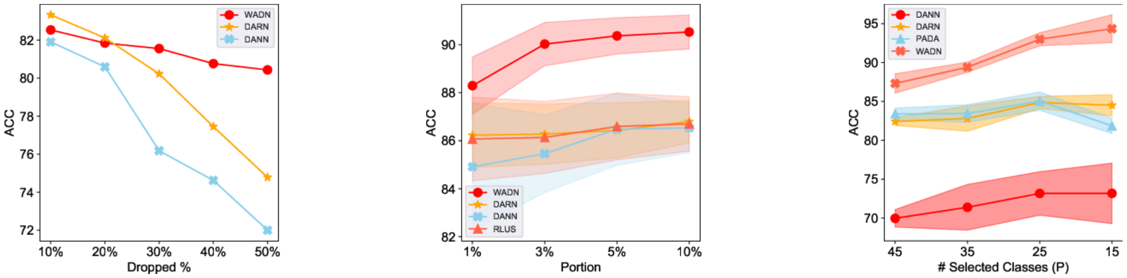
The empirical results reveal a significantly better performance ($\approx 2\% - 6\%$) on different benchmarks. For understanding the aggregation principles of WADN, we visualize the task relations in digits (Fig. 4(a)) with demonstrating a *non-uniform* $\boldsymbol{\lambda}$, which highlights the importance of properly choosing the most related source rather than simply merging all the data. For example, when the target domain is SVHN, WADN mainly leverages the information from SYNTH, since they are more semantically similar, and MNIST does not help too much for SVHN, which is also observed by (Ganin et al., 2016). Besides, Fig. 4(b) visualizes the evolution of $\boldsymbol{\lambda}$ between WADN and recent principled approach DARN (Wen et al., 2020), which utilized the $\mathbb{P}(x)$ information and dynamic updating to find the similar domains. Compared with WADN, $\boldsymbol{\lambda}$ in DARN is *unstable* during updating under drifted label distribution.

Besides, we conduct the ablation study through evaluating the performance under different levels of source label shift in Amazon Review dataset (Fig. 5(a)). The results show strong practical benefits for WADN in the larger label shift. The additional analysis and results can be found in Appendix.

6.2. Multi-Source DA with Limited Target Labels

We adopt Amazon Review and Digits in the multi-source DA with limited target samples, which have been widely used. In the experiments, we still use shifted sources. We randomly sample only 10% labeled samples (w.r.t. target dataset in unsupervised DA) as training set and the rest 90% samples as the unseen target test set. We adopt the same hyper-parameters and training strategies with unsupervised DA. We specifically add two recent baselines RLUS (Konstantinov & Lampert, 2019) and MME (Saito et al., 2019), which also considered DA with the labeled target domain.

The results are reported in Tab. 3, which also indicates strong empirical improvement. Interestingly, on the Amazon review dataset, the previous aggregation approach RLUS is unable to select the related source when label distribution varies. To show the effectiveness of WADN, we test various portions of labelled samples ($1\% \sim 10\%$) on the target. The results in Fig. 5(b) on USPS dataset show consistently better than the baseline, even in the few target samples scenarios



(a) (b) (c)

Figure 5. Ablation study on different scenarios. (a) Unsupervised DA with Amazon Review dataset. Accuracy under different levels of label shifted sources (higher dropping rate means larger label drift). The results are reported on the average of all the domains, see the results for each domain in Appendix. (b) Multi-Source DA with limited target label in digits task with target USPS. The performance (mean ± std) of WADN is consistently better under different target samples (smaller portion indicates fewer target samples). (C) Partial Multi-source DA in office-home dataset with target domain Product. Performance (mean ± std) of different number of selected classes on the target, where WADN shows a consistent better performance under different selected sub-classes.

Table 4. Unsupervised Multi-Source Partial DA: Accuracy $(\%)$ on Office-Home (#Source: 65, #Target: 35)

Target	Art	Clipart	Product	Real-World	Average
Source	$50.56_{\pm 1.42}$	$49.79_{\pm 1.14}$	$68.10_{\pm 1.33}$	$78.24_{\pm 0.76}$	61.67
DANN	$53.86_{\pm 2.23}$	$52.71_{\pm 2.20}$	$71.25_{\pm 2.44}$	$76.92_{\pm 1.21}$	63.69
MDAN	$67.56_{\pm 1.39}$	$65.38_{\pm 1.30}$	$81.49_{\pm 1.92}$	$83.44_{\pm 1.01}$	74.47
MDMN	$68.13_{\pm 1.08}$	$65.27_{\pm 1.93}$	$81.33_{\pm 1.29}$	$84.00_{\pm 0.64}$	74.68
M^{3}SDA	$65.10_{\pm 1.97}$	$61.80_{\pm 1.99}$	$76.19_{\pm 2.44}$	$79.14_{\pm 1.51}$	70.56
DARN	$71.53_{\pm 0.63}$	$69.31_{\pm 1.08}$	$82.87_{\pm 1.56}$	$84.76_{\pm 0.57}$	77.12
PADA	$74.37_{\pm 0.84}$	$69.64_{\pm 0.80}$	$83.45_{\pm 1.13}$	$85.64_{\pm 0.39}$	78.28
WADN	$\mathbf{80.06}_{\pm 0.93}$	$\mathbf{75.90}_{\pm 1.06}$	$\mathbf{89.55}_{\pm 0.72}$	$\mathbf{90.40}_{\pm 0.39}$	**83.98**

such as $1 - 3\%$.

6.3. Partial Unsupervised Multi-Source DA

In this scenario, we adopt the Office-Home dataset to evaluate our approach, as it contains large (65) classes. We do not change the source domains and we randomly choose 35 classes from the target. We evaluate all the baselines on the same selected classes and repeat 5 times. All reported results are averaged from 3 different sub-class selections (15 runs in total), shown in Tab. 4. We additionally compare PADA (Cao et al., 2018) approach by merging all sources and use one-to-one partial DA algorithm. We adopt the same hyper-parameters and training strategies in unsupervised DA scenario.

The reported results are also significantly better than the current multi-source DA or one-to-one partial DA approach, which again emphasizes the benefits of WADN: properly selecting the related sources by using semantic information.

Besides, we change the number of selected classes (Fig 5(c)), the proposed WADN still indicates consistent better results by a large margin, which indicates the importance of considering $\hat{\alpha}_t$ and $\boldsymbol{\lambda}$. In contrast, DANN shows unstable results on average in less selected classes. Beside, WADN shows a good estimation of the label distribution ratio (Fig 6) and has correctly detected the non-overlapping classes, which verifies the effectiveness of the label-distribution estimator and indicates its good explainability.

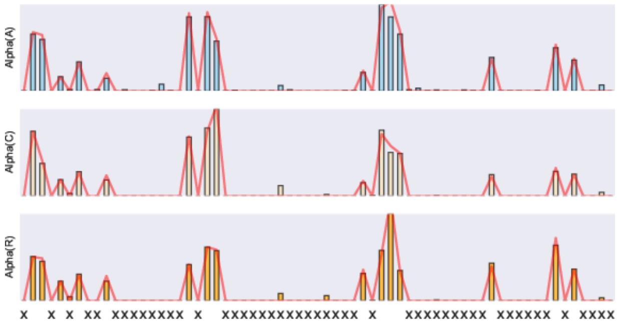


Figure 6. Analysis on Partial DA of target Product. We select 15 classes and visualize estimated $\hat{\alpha}_t$ (the bar plot). The "X" along the x-axis represents the index of *dropped* 50 classes. The red curves are the true label distribution ratio. See Appendix for additional results and analysis.

7. Conclusion

In this paper, we proposed a novel algorithm WADN for multi-source domain adaptation problem under different label proportions. WADN differs from previous approaches in two key prospects: a better source aggregation approach when label distributions change; a unified empirical framework for three popular DA scenarios. We evaluated the proposed method by extensive experiments and showed its strong empirical results.

Acknowledgments

C. Shui and C. Gagné acknowledge support from NSERC-Canada and CIFAR. B. Wang is supported by NSERC Discovery Grants Program.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl0r3R9KX.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pp. 137–144, 2007.

Ben-David, S., Lu, T., Luu, T., and Pál, D. Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pp. 129–136, 2010.

Blitzer, J., Dredze, M., and Pereira, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 440–447, 2007.

Cao, Z., Ma, L., Long, M., and Wang, J. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 135–150, 2018.

Chapelle, O. and Zien, A. Semi-supervised classification by low density separation. In *AISTATS*, volume 2005, pp. 57–64. Citeseer, 2005.

Chen, M., Xu, Z., Weinberger, K. Q., and Sha, F. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1627–1634, 2012.

Combes, R. T. d., Zhao, H., Wang, Y.-X., and Gordon, G. Domain adaptation with conditional distribution matching and generalized label shift. *arXiv preprint arXiv:2003.04475*, 2020.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. C. A unified view of label shift estimation. *arXiv preprint arXiv:2003.07554*, 2020.

Geiss, L. S., Wang, J., Cheng, Y. J., Thompson, T. J., Barker, L., Li, Y., Albright, A. L., and Gregg, E. W. Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, united states, 1980-2012. *Jama*, 312 (12):1218–1226, 2014.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pp. 2839–2848, 2016.

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

Hoffman, J., Mohri, M., and Zhang, N. Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems*, pp. 8246–8256, 2018a.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., and Darrell, T. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pp. 1989–1998. PMLR, 2018b.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for nlp. *arXiv preprint arXiv:1902.00751*, 2019.

Hull, J. J. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

Johansson, F., Sontag, D., and Ranganath, R. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 527–536, 2019.

Konstantinov, N. and Lampert, C. Robust learning from untrusted sources. In *International Conference on Machine Learning*, pp. 3488–3498, 2019.

Li, H., Jialin Pan, S., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, 2018a.

Li, Y., Carlson, D. E., et al. Extracting relationships by multi-domain matching. In *Advances in Neural Information Processing Systems*, pp. 6798–6809, 2018b.

Li, Y., Murias, M., Major, S., Dawson, G., and Carlson, D. On target shift in adversarial domain adaptation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 616–625, 2019.

Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pp. 3122–3130, 2018.

Liu, J., Hong, Y., D'Agostino Sr, R. B., Wu, Z., Wang, W., Sun, J., Wilson, P. W., Kannel, W. B., and Zhao, D. Predictive value for the chinese population of the framingham chd risk assessment tool compared with the chinese multi-provincial cohort study. *Jama*, 291(21): 2591–2599, 2004.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Mansour, Y., Mohri, M., Suresh, A. T., and Wu, K. A theory of multiple-source adaptation with limited target labeled data. *arXiv preprint arXiv:2007.09762*, 2020.

Mohri, M. and Medina, A. M. New analysis and algorithm for learning with drifting distributions. In *International Conference on Algorithmic Learning Theory*, pp. 124–138. Springer, 2012.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2009.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems*, pp. 3347–3357, 2019.

Redko, I., Courty, N., Flamary, R., and Tuia, D. Optimal transport for multi-source domain adaptation under target shift. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 849–858. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/redko19a.html.

Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pp. 15–18, 2019.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732, 2018.

Saito, K., Kim, D., Sclaroff, S., Darrell, T., and Saenko, K. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8050–8058, 2019.

Shui, C., Abbasi, M., Robitaille, L.-É., Wang, B., and Gagné, C. A principled approach for learning task similarity in multitask learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 3446–3452, 2019.

Sugiyama, M. and Kawanabe, M. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5018–5027, 2017.

Wang, B., Mendez, J., Cai, M., and Eaton, E. Transfer learning via minimizing the performance gap between domains. In *Advances in Neural Information Processing Systems*, pp. 10645–10655, 2019a.

Wang, B., Zhang, H., Liu, P., Shen, Z., and Pineau, J. Multitask metric learning: Theory and algorithm. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 3362–3371, 2019b.

Wang, B., Wong, C. M., Kang, Z., Liu, F., Shui, C., Wan, F., and Chen, C. P. Common spatial pattern reformulated for regularizations in brain-computer interfaces. *IEEE Transactions on Cybernetics*, 2020.

Wen, J., Greiner, R., and Schuurmans, D. Domain aggregation networks for multi-source domain adaptation. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Wu, Y., Winston, E., Kaushik, D., and Lipton, Z. Domain adaptation with asymmetrically-relaxed distribution alignment. In *International Conference on Machine Learning*, pp. 6872–6881, 2019.

Zhang, J., Li, W., Ogunbona, P., and Xu, D. Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 2019.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pp. 819–827, 2013.

Zhao, H., Zhang, S., Wu, G., Moura, J. M., Costeira, J. P., and Gordon, G. J. Adversarial multiple source domain adaptation. In *Advances in neural information processing systems*, pp. 8559–8570, 2018.

Zhao, S., Li, B., Xu, P., and Keutzer, K. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*, 2020.