This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

COT: Unsupervised Domain Adaptation with Clustering and Optimal Transport

Yang Liu [1] * Zhipeng Zhou [1] * Baigui Sun [1] †
[1]Alibaba Group

Abstract

Unsupervised domain adaptation (UDA) aims to transfer the knowledge from a labeled source domain to an unlabeled target domain. Typically, to guarantee desirable knowledge transfer, aligning the distribution between source and target domain from a global perspective is widely adopted in UDA. Recent researchers further point out the importance of local-level alignment and propose to construct instance-pair alignment by leveraging on Optimal Transport (OT) theory. However, existing OT-based UDA approaches are limited to handling class imbalance challenges and introduce a heavy computation overhead when considering a large-scale training situation. To cope with two aforementioned issues, we propose a Clustering-based Optimal Transport (COT) algorithm, which formulates the alignment procedure as an Optimal Transport problem and constructs a mapping between clustering centers in the source and target domain via an end-to-end manner. With this alignment on clustering centers, our COT eliminates the negative effect caused by class imbalance and reduces the computation cost simultaneously. Empirically, our COT achieves state-of-the-art performance on several authoritative benchmark datasets.

1. Introduction

Benefiting from the availability of large-scale data, deep learning has achieved tremendous success over the past few years. However, directly applying a well-trained convolution neural network on a new domain frequently suffers from the domain gap/discrepancy challenge, resulting in spurious predictions on the new domain. To remedy this, Unsupervised Domain Adaptation (UDA) has attracted many researchers' attention, which can transfer the knowledge from a labeled domain to an unlabeled domain.

A major line of UDA approaches [1, 1, 28, 42, 49, 53] aim to learn a global domain shift by aligning the global source and target distribution while ignoring the local-level alignment between two domains. By leveraging on global domain adaptation, the global distributions of source and target domain are almost the same, thus losing the fine-grained information for each class (class-structure) on the source and target domain.

Recently, to preserve class structure in both domains, several works [6, 15, 23, 30, 38, 40, 44, 51, 54] adopt optimal transport (OT) to minimize the sample-level transportation cost between source and target domain, achieving a significant performance on UDA. However, there exist two issues on recent OT-based UDA approaches. (i) When considering a realistic situation, i.e. the class imbalance[1] phenomenon occurs between the source and target domain, samples belonging to the same class in the target domain are assigned with false pseudo labels due to the mechanism of optimal transport, which requires each sample in source domain can be mapped to target samples. As a result, existing OT-based UDA methods provide poor pair-wise matching when facing class imbalance challenges. (ii) OT-based UDA methods tend to find a sample-level optimal counterpart, which requires a large amount of computation overhead, especially training on large-scale datasets.

To solve two aforementioned issues, we propose a Clustering-based Optimal Transport algorithm, termed COT, to construct a clustering-level instead of sample-level mapping between source and target domain. Clusters in the source domain are obtained from the classifiers supervised by the labeled source domain data. While for the target domain, COT utilizes a set of learnable clusters to represent the feature distribution of the target domain, which can describe the sub-domain information [50, 57]. For instance, in many object recognition tasks [13, 20] an object could contain many attributes. Each attribute can be viewed as a sub-domain. To this end, the clusters on the source and target domain can represent the individual sub-domain information, respectively, such that optimal transport between clusters intrinsically provides a local mapping from the sub-domain in the source domain to those in the target domain. Moreover, we provide a theoretical analysis and compre-

Email: ly261666@alibaba-inc.com
* Equal Contribution
† Corresponding Author

[1]label distribution are different in two domains, $\boldsymbol{P}_s(y) \neq \boldsymbol{P}_t(y)$

hensive experimental results to guarantee that (i) COT can alleviate the negative effect caused by class imbalance; (ii) Compared to existing OT-based UDA approaches, our COT economizes much computation head.

In summary, our main contributions include:

- We propose a novel Clustering-based Optimal Transport module as well as a specially designed loss derived from the discrete type of Kantorovich dual form, which resolves two aforementioned challenges on the existing OT-based UDA algorithms, facilitating the development of OT-based UDA community.
- We provide a theoretical analysis to guarantee the advantages of our COT.
- Our COT achieves state-of-the-art performance on several UDA benchmark datasets.

2. Related work

Pseudo Label based Domain Adaptation Inspired by the observation that samples in the target domain can be clustered within the feature space, for accurate pseudo-labeling, [48] propose a selective pseudo-labeling strategy based on structural predictions which utilize the unsupervised clustering analysis. [41] introduce a confidence-based weighting scheme for obtaining pseudo-labels and an adaptive threshold adjustment strategy to provide sufficient and accurate pseudo-labels during the training process. The confidence-based weighting scheme generates pseudo-labels that can enable the performance less sensitive to the threshold determined by the pseudo-labels. In the task of person re-identification, [17] propose an unsupervised framework called Mutual Mean-Teaching to learn better features from the target domain by refining the hard pseudo labels offline and soft pseudo labels online alternatively to mitigate the effects of noisy pseudo labels caused by the clustering algorithms.

Optimal Transport based Domain Adaptation As a way to find a minimal effort strategy to the transport of a given mass of dirt into a given hole, [37] put forward the optimal transport problem for the first time. [22] provide an extension of the original problem of Monge. Recently, by applying the optimal transport to domain adaptation, some new computation strategies have emerged.

[7] propose regularized unsupervised optimal transport model to align the representation of features between different domains. The regularization schemes encoding class-structure in source domain during estimation of transport map enforce the intuition that samples of the same class must undergo a similar transformation. [5] minimize the optimal transport loss between the joint source distribution and the estimated target joint distribution depending on a function that is introduced to predict an output value given input from the source domain. For reducing the discrepancy between multiple domains, [40] propose Joint Class Proportion and Optimal Transport which performs multi-source domain adaptation and target shift correction simultaneously by learning the predicted class probability of the unlabeled target data and the coupling to align the distributions between source and target domain. For better alignment between different domains, a relation between target error and the magnitude of different Wasserstein distances are proposed in [23] which optimize the metric for domain adaptation.

3. Preliminary

In this section, we introduce the basic knowledge for optimal transport.

3.1. Optimal Transport

Let $X \subseteq \mathbb{R}^d$ be a measurable space and the labels are denoted as $\mathcal{Y}$. We denote the set of all probability distributions on X as $\mathcal{P}(X)$. The source and target domains are space X equipped with two distinct probability distributions $\mu_{\mathcal{S}}$ and $\mu_{\mathcal{T}}$. Suppose we have source dataset $\{x_i^s\}_{i=1}^{n_s} \subset X_{\mathcal{S}} = (X, \mu_{\mathcal{S}})$ associated with label set $\{y_i^s\}_{i=1}^{n_s}$ where $y_i^s \in \mathcal{Y}$. The target dataset is $\{x_j^t\}_{j=1}^{n_t} \subset X_{\mathcal{T}} = (X, \mu_{\mathcal{T}})$ without labels. The goal of optimal transport is to minimize the inter-domain transportation cost by finding a feasible map to preserve measure.

Definition 1 *(Kantorovich) For given joint distribution $\rho(x^s, x^t)$ which satisfies for every measurable Borel set $O_{\mathcal{S}} \subset X_{\mathcal{S}}, O_{\mathcal{T}} \subset X_{\mathcal{T}}$, we have*

$$\rho(O_{\mathcal{S}} \times X_{\mathcal{T}}) = \mu_{\mathcal{S}}(O_{\mathcal{S}}), \rho(X_{\mathcal{S}} \times O_{\mathcal{T}}) = \mu_{\mathcal{T}}(O_{\mathcal{T}}) \quad (1)$$

For convenience, we denote the projection maps from $X_{\mathcal{S}} \times X_{\mathcal{T}}$ to $X_{\mathcal{S}}$ and $X_{\mathcal{T}}$ as $\pi_{\mathcal{S}}, \pi_{\mathcal{T}}$. The above equation can be denoted as $\pi_{\mathcal{S}\#}\rho = \mu_{\mathcal{S}}$ and $\pi_{\mathcal{T}\#}\rho = \mu_{\mathcal{T}}$. The corresponding transportation cost is

$$\mathcal{C}(\rho) = \int_{X_{\mathcal{S}} \times X_{\mathcal{T}}} c(x^s, x^t) d\rho(x^s, x^t) \quad (2)$$

where $c(x^s, x^t)$ is pointwise transportation cost between $x^s \in X_{\mathcal{S}}$ and $x^t \in X_{\mathcal{T}}$. The optimal transport problem is proposed to minimize the $\mathcal{C}(\rho)$ under the measure preserving as the following:

$$W_c := \inf_{\rho} \{\mathcal{C}(\rho) || \pi_{\mathcal{S}\#}\rho = \mu_{\mathcal{S}}, \pi_{\mathcal{T}\#}\rho = \mu_{\mathcal{T}}\} \quad (3)$$

By convex optimization theory, we can consider the Kantorovich's dual problem as:

$$W_c := \begin{cases} \max\limits_{\varphi,\psi} \int_{X_{\mathcal{S}}} \varphi(x^s) d\mu_{\mathcal{S}} + \int_{X_{\mathcal{T}}} \psi(x^t) d\mu_{\mathcal{T}} \\ \text{s.t.} \quad \varphi(x^s) + \psi(x^t) \leq c(x^s, x^t) \end{cases} \quad (4)$$

where φ and ψ are real functions from $X_{\mathcal{S}}$ and $X_{\mathcal{T}}$ to $\mathbb{R}$. Moreover, the Kantorovich problem can be formulated as

$$\max_{\varphi}\{\int_{X_{\mathcal{S}}}\varphi(x^s)d\mu_{\mathcal{S}}+\int_{X_{\mathcal{T}}}\varphi^c(x^t)d\mu_{\mathcal{T}}\} \tag{5}$$

where $\varphi^c(x^t)=\inf_{x_s\in X_{\mathcal{S}}}\{c(x^s,x^t)-\varphi(x^s)\}$ is called the c-transform of φ.

By classical optimal transport theory, different choices of cost function will influence the difficulty to solving the optimal transport problem. When we choose $c(x^s,x^t)=\|x^s-x^t\|_2$, the problem stated in Equation (5) is equivalent to

$$\max_{\varphi}\{\int_{X_{\mathcal{S}}}\varphi(x^s)d\mu_{\mathcal{S}}-\int_{X_{\mathcal{T}}}\varphi(x^t)d\mu_{\mathcal{T}}\} \tag{6}$$

where φ is under the constraint that $|\varphi(x)-\varphi(x')|\leq\|x-x'\|_2$. WGAN [2] is inspired by above cost setting, during the implementation of optimal transport in WGAN, they utilize the gradient clip to guarantee the Lipschitz constant of φ is bounded from above by 1. When we set the cost function as $c(x^s,x^t)=\|x^s-x^t\|_2^2$, [16] guarantees the existence and uniqueness of optimal transport map.

4. Methodology

In this section, we present the detail of our COT and demonstrate the advantages of our COT theoretically.

4.1. Clustering-based Optimal Transport

Instead of aligning instance-level features between source and target domain, we propose a novel clustering-based optimal transport (COT) module for unsupervised domain adaptation in this subsection. Firstly, we extract features from the source and target domain by ImageNet pre-trained CNNs. Then we utilize learnable clusters to represent the sub-domains in the source and target domain, respectively. Finally, we apply a Kantorovich dual form-based loss to implement the optimal transport between clusters from both domains.

Feature Extractor We utilize an ImageNet pre-trained (without fully connected layers) CNNs (e.g. ResNet50/ResNet101) to extract features $\{x_i^s\}_{i=1}^{n_s}$ and $\{x_j^t\}_{j=1}^{n_t}$ from the source and target dataset respectively at the beginning of training process. Note that the distributions of features vary during the training phase.

Clustering As for each sample from source domain, i.e., x_i^s and corresponding ground-truth $y_i^s\in\mathcal{Y}$. We denote the fully-connected layer which outputs the classification logits as $W=[w_1^s,...,w_{|\mathcal{Y}|}^s]^\top\in\mathbb{R}^{|\mathcal{Y}|\times c}$, where $|\mathcal{Y}|$ is the number of categories and c is number of feature channels. The predicted classification probability is $P(\hat{y}_i^s=v|x_i^s)=\frac{e^{x_i^{s\top}w_v^s}}{\sum_{u=1}^{|\mathcal{Y}|}e^{x_i^{s\top}w_u^s}}$. The corresponding cross-entropy loss is shown as follows:

$$\mathcal{L}_{cross-entropy}=\frac{1}{b}\sum_{i=1}^{b}-y_i^s\cdot\log(P(\hat{y}_i^s|x_i^s)) \tag{7}$$

For the source domain, since $\{w_v^s\}_{v=1}^{|\mathcal{Y}|}$ [13, 47] have been shown effective for clustering representation, we take the classifiers $\{w_v^s\}_{v=1}^{|\mathcal{Y}|}$ as clusters for feature space of source domain. For the target domain, a set of learnable clusters termed as $\{w_u^t\}_{u=1}^{K}$ are proposed to represent the sub-domains, where $K=Q\cdot|\mathcal{Y}|$ is a hyper-parameter which stands for the number of sub-domains in the target domain, Q is a positive integer which represents the number of sub-domains for each class. The motivation of introducing Q is to preserve the sub-domain information in the target domain during the COT optimization procedure. For instance, if we set the same number of sub-domains on the source and target domain, the sub-domain information in the target domain would be seriously influenced by those in the source domain during the COT procedure, since the latter is optimized with hard supervision signal while the former is supervised by soft signal. On the contrary, Q helps the target domain generate more sub-domains, which can preserve the sub-domain information in the target domain because not all of them are optimized during the COT phase. Empirically, the performance is superior when Q is set to 2. $\{w_u^t\in\mathbb{R}^c\}_{u=(l-1)\cdot Q+1}^{l\cdot Q}$ represent the clusters for class l, $\forall 1\leq l\leq|\mathcal{Y}|$. For each feature x_j^t, we assign it to the closest cluster in $\{w_u^t\}_{u=1}^{K}$. We utilize the L_2 distance to measure the distance between features and clusters and pull the features back to corresponding clusters.

$$\begin{cases}\mathcal{L}_{\text{cluster}}=\sum_{j=1}^{n_t}\|w_{u*}^t-x_j^t\|^2\\ s.t.\ \ u_j^*=\arg\min_{1\leq u\leq K}(\|w_u^t-x_j^t\|^2),\forall 1\leq j\leq n_t\end{cases} \tag{8}$$

Optimal Transport With clusters $\{w_v^s\}_{v=1}^{|\mathcal{Y}|}$ and $\{w_u^t\}_{u=1}^{K}$ from source and target domain respectively, we design the clustering based optimal transport as follows:

$$\begin{cases}\min_{T\in\mathbb{R}^{|\mathcal{Y}|\times K}}T_{vu}c_{vu}\\ s.t.\sum_{u=1}^{K}T_{vu}=\frac{1}{K},\sum_{v=1}^{|\mathcal{Y}|}T_{vu}=\frac{1}{|\mathcal{Y}|},\\ \ \ T_{vu}\geq 0,\forall 1\leq v\leq|\mathcal{Y}|,1\leq u\leq K\end{cases} \tag{9}$$

where $c_{vu}=\|w_v^s-w_u^t\|_2^2$. Similar to Equation (5), we can get the discrete Kantorovich dual problem of Equation (9).

$$\max_{\psi}\{\frac{1}{K}\sum_{u=1}^{K}\psi(w_u^t)+\frac{1}{|\mathcal{Y}|}\sum_{v=1}^{|\mathcal{Y}|}\psi^c(w_v^s)\} \tag{10}$$

Algorithm 1 Clustering-based Optimal Transport

1: Set number of epochs for training as E, learnable clusters for target domain as $\{w_u^t\}_{u=1}^{|\mathcal{Y}|}$, classifiers/clusters for source domain as $\{w_v^s\}_{v=1}^{|\mathcal{Y}|}$;
2: **for** k-th training epoch while $k \leq E$ **do**
3: **for** t-th iteration in k-th epoch **do**
4: Take mini-batch of samples from source and target domain as input for feature extractor CNNs with parameters θ, the output features are $\{x_i^s\}_{i=1}^b$ and $\{x_j^t\}_{j=1}^b$;
5: Compute the L_{cluster} for $\{x_j^t\}_{j=1}^b$, $L_{\text{cross-entropy}}$ for $\{x_i^s\}_{i=1}^b$ in the l-th batch, and $\mathcal{L}_{OT}$;
6: **if** $1 \leq t \leq k/(b*10)$ **then**
7: we find the current optimal map from clusters of source domain to those of target domain by maximizing L_{OT};
8: **end if**
9: Minimize $\mathcal{L}_{COT}$;
10: **end for**
11: **end for**

where $\psi^c(w_v^s) = \inf_{u=1}^{K}(c_{vu} - \psi(w_u^t))$. According to Equation 10, We can seek for the optimal transportation map between clusters by optimizing the following loss (line 7 in Algorithm 1):

$$\mathcal{L}_{\text{OT}} = \frac{1}{K}\sum_{u=1}^{K}\lambda_u^t + \frac{1}{|\mathcal{Y}|}\sum_{v=1}^{|\mathcal{Y}|}(\inf_{u=1}^{K}(c_{vu} - \lambda_u^t)) \tag{11}$$

where $\{\lambda_u^t\}_{u=1}^K$ represent the value of function ψ at points $\{w_u^t\}_{u=1}^K$. Furthermore, it is worth noting that cost c_{vu} is frozen during the optimization of $\mathcal{L}_{\text{OT}}$, indicating that only $\{\lambda_v^t\}_{v=1}^{|\mathcal{Y}|}$ is updated in this step. Then based on the founded optimal map, we minimize the following loss ($\mathcal{L}_{\text{COT}}$) to close the matched clusters from the feature representation perspective, thus the domain-invariant feature representation is learned via this step (line 9 in Algorithm 1). Correspondingly, we freeze the $\{\lambda_v^t\}_{v=1}^{|\mathcal{Y}|}$ and only update c_{vu} in this step.

$$\mathcal{L}_{\text{COT}} = \mathcal{L}_{\text{cross-entropy}} + \alpha_1 \cdot \mathcal{L}_{\text{cluster}} + \alpha_2 \cdot \mathcal{L}_{\text{OT}} \tag{12}$$

where α_1 and α_2 are loss weights. Finally, we elaborate the overall optimization strategy of our COT in Algorithm 1. Line 6 of Algorithm 1 represents we only optimize $\mathcal{L}_{OT}$ in the initial 10% iterations of each epoch.

4.2. Theoretical Analysis on Instance/Clustering Optimal Transport

Given features $\{x_i^s\}_{i=1}^{n_s}$ and $\{x_j^t\}_{j=1}^{n_t}$ from source and target domain respectively, where x_i^s and x_j^t are output from shared-parameters neural network for feature extractor. We consider the discrete Kantorovich problem

$$\begin{cases} \min\limits_{T\in\mathbb{R}^{n_s\times n_t}} T_{ij}c_{ij} \\ s.t. \sum\limits_{j=1}^{n_t} T_{ij} = \frac{1}{n_s}, \sum\limits_{i=1}^{n_s} T_{ij} = \frac{1}{n_t}, \\ T_{ij} \geq 0, \forall 1 \leq i \leq n_s, 1 \leq j \leq n_t. \end{cases} \tag{13}$$

where $c_{ij} = \|x_i^s - x_j^t\|_2^2$.

Considering the distance and inner-product between features and classifiers:

$$\begin{aligned} &\|x_i^s - w_{v_1^s}\|_2^2 - \|x_i^s - w_{v_2}^s\|_2^2 \\ =&\|w_{v_1}^s\|_2^2 - \|w_{v_2}^s\|_2^2 \\ +&2(\|w_{v_2}^s\|_2 x_i^{s\top}\frac{w_{v_2}^s}{\|w_{v_2}^s\|_2} - \|w_{v_1}^s\|_2 x_i^{s\top}\frac{w_{v_1}^s}{\|w_{v_1}^s\|_2}) \end{aligned} \tag{14}$$

In the Bayesian view, we can consider $\|w_v^s\|_2$ as the prior probability of class v, x_i^s is feature representation of a sample and $\frac{w_v^s}{\|w_v^s\|_2}$ is the cluster for class v. $x_i^{s\top}\frac{w_v^s}{\|w_v^s\|_2}$ measure the similarity between feature and cluster. When classifiers in $\{\|w_v^s\|_2\}_{v=1}^{|\mathcal{Y}|}$ are of the same magnitude, we conclude that the similarity between features and clusters are almost equivalent to the distance between features and classifiers. With labels as supervision, the optimization of cross-entropy can promote the inter-class discrepancy which implies

$$x_i^{s\top}\frac{w_{y_i^s}^s}{\|w_{y_i^s}^s\|_2} \gg x_i^{s\top}\frac{w_v^s}{\|w_v^s\|_2}, \forall v \neq y_i^s \tag{15}$$

which also provide the following result

$$\|x_i^s - w_{y_i^s}^s\|_2^2 \ll \|x_i^s - w_v^s\|_2^2, \forall v \neq y_i^s \tag{16}$$

If clustering doesn't work sufficiently well, it happens that some samples in the source domain with label v are assigned to samples in the target domain with label $u \neq v$. When clustering performs well, We have $c_{ij} \sim \bar{c}_{vj} = \|w_v^s - x_j^t\|_2^2$, where $\sim$ means these two numbers are almost the same. Then we get

$$\sum_{i,j} T_{ij}c_{ij} \sim \sum_{v,j}(\sum_{x_i^s\in X_v^s} T_{ij})\bar{c}_{vj} \tag{17}$$

where X_v^s is the set of samples with label v in the source domain. We denote the number of samples with class v as n_v, then we get

$$\sum_{j=1}^{n_t}(\sum_{x_i^s\in X_v^s} T_{ij}) = \frac{n_v}{n_s}, \sum_{v=1}^{|\mathcal{Y}|}(\sum_{x_i^s\in X_v^s} T_{ij}) = \frac{1}{n_t} \tag{18}$$

Table 156 Accuracy (%) on Office-31 for UDA (ResNet-50).The best result is in bold.

	Method	A→W	D→W	W→D	A→D	D→A	W→A	Avg
Common UDA	ADDA [46]	86.2	96.2	98.4	77.8	69.5	68.9	82.9
	JAN [34]	85.4	97.4	99.8	84.7	68.6	70.0	84.3
	MCD [42]	88.6	98.5	**100.0**	92.2	69.5	69.7	86.5
	GTA [43]	89.5	97.9	99.8	87.7	72.8	71.4	86.5
	CDAN [33]	94.1	98.6	**100.0**	92.9	71.0	69.3	87.7
	TAT [32]	92.5	99.3	**100.0**	93.2	73.1	72.1	88.4
	MDD [25]	94.5	98.4	**100.0**	93.5	74.6	72.2	88.9
	GSP [19]	92.9	98.7	99.8	94.5	75.9	74.9	89.5
	DANN [1]	82.0	96.9	99.1	79.7	68.2	67.4	82.2
	SHOT [29]	94.0	90.1	74.7	98.4	74.3	99.9	88.6
	MCC [21]	95.5	98.6	**100.0**	94.4	72.9	74.9	89.4
	GVB-GD [9]	94.8	98.7	**100.0**	95.0	73.4	73.7	89.3
	TSA [28]	96.0	98.7	**100.0**	95.4	**76.7**	76.8	90.6
	SRDC [45]	95.7	**99.2**	**100.0**	95.8	**76.7**	77.1	90.8
OT-based	JDOT [5]	84.7	97.8	**100.0**	86.4	64.4	67.7	83.5
	DeepJDOT [11]	88.9	98.5	99.6	88.2	72.1	70.1	86.2
	MLOT [23]	92.8	98.5	**100.0**	90.8	72.8	71.6	87.8
	RWOT [51]	95.1	99.5	**100.0**	94.5	77.5	77.9	90.8
	DANN [1] + MMI [28]	95.2	98.6	100.0	94.4	74.6	75.2	89.7
	DANN + MMI + COT (Ours)	**96.5**	99.1	**100.0**	**96.1**	**76.7**	**77.4**	**91.0**

Inspired by Equation (17) and (18), we consider the following optimal transport between clusters from the source domain and instances from the target domain instead of solving the Kantorovich problem in Equation (13).

$$\begin{cases} \min\limits_{\bar{T}\in\mathbb{R}^{|\mathcal{Y}|\times n_t}} \bar{T}_{vj}\bar{c}_{vj} \\ s.t. \sum\limits_{j=1}^{n_t}\bar{T}_{vj} = \dfrac{n_v}{n_s}, \sum\limits_{v=1}^{|\mathcal{Y}|}\bar{T}_{vj} = \dfrac{1}{n_t}, \\ \bar{T}_{vj} \geq 0, \forall 1 \leq v \leq |\mathcal{Y}|, 1 \leq j \leq n_t \end{cases} \tag{19}$$

In general, because of the class imbalance, the empirical label distribution between source and target dataset are different

$$\exists \delta > 0, \text{ s.t. } \|(\frac{n_1^s}{n_s}, ..., \frac{n_{|\mathcal{Y}|}^s}{n_s}) - (\frac{n_1^t}{n_t}, ..., \frac{n_{|\mathcal{Y}|}^t}{n_t})\|_2 \geq \delta \tag{20}$$

where δ is a constant which measures the label distribution between source and target domain. There must exists some index i such that $\frac{n_v^s}{n_s} > \frac{n_v^t}{n_t}$, which means that some samples with label v in the source domain will be assigned to samples in the target domain with label $u \neq v$. This will result in samples belonging to the same category in the target domain are given different pseudo labels, which increase the difficulty of training and cause the degradation of the performance of deep learning methods on the target domain.

When we utilize the clustering based optimal transport, for source domain, we have $\sum\limits_{u=1}^{K} T_{vu} = \frac{1}{|\mathcal{Y}|}$. For target domain, $\sum\limits_{v=1}^{|\mathcal{Y}|}\sum\limits_{u=(l-1)\cdot Q+1}^{l\cdot Q} T_{vu} = \frac{Q}{K} = \frac{1}{|\mathcal{Y}|}$, which ease the negative effect from class imbalance in domain adaptation based on optimal transport.

4.3. Computation Cost

In terms of instance-based optimal transport, firstly we need to obtain the features of all samples from the source and target domain, computation cost on the feature extractor is shown as follows:

$$\mathcal{O}(n_s + n_t) \cdot \mathcal{O}(\text{feature-extractor}) \tag{21}$$

where $\mathcal{O}(\text{feature-extractor})$ means the computation cost on single sample when extracting the feature. Then considering the optimization of optimal transport, every iteration will need $\mathcal{O}(n_s \cdot n_t)$. In comparison, for cluster-based optimal transport, the main computation cost on optimal transport is $\mathcal{O}(|\mathcal{Y}| \cdot K)$. For a large-scale dataset, the clustering-based optimal transport cost is much less than instance-based optimal transport.

5. Experiments

In this section, we first elaborate on implementation details and 3 authoritative benchmark datasets in the field of UDA. Then we compare our COT with existing OT-based UDA algorithms and state-of-the-art UDA methods, illustrating that our COT achieves the dominant result in the field of OT-based UDA and competitive performance on the common UDA realm, respectively. Furthermore, we provide some qualitative analysis to illustrate the advantage of our COT when compared with some clustering-based DA

Table 256 Accuracy (%) on Office-Home for UDA.The best result is in bold.

	Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
Common UDA	JAN [34]	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
	TAT [32]	51.6	69.5	75.4	59.4	69.5	68.6	59.5	50.5	76.8	70.9	56.6	81.6	65.8
	TPN [39]	51.2	71.2	76.0	65.1	72.9	72.8	55.4	48.9	76.5	70.9	53.4	80.4	66.2
	ETD [26]	51.3	71.9	**85.7**	57.6	69.2	73.7	57.8	51.2	79.3	70.2	57.5	82.1	67.3
	SymNets [56]	47.7	72.9	78.5	64.2	71.3	74.2	64.2	48.8	79.5	74.5	52.6	82.7	67.6
	BNM [8]	52.3	73.9	80.0	63.3	72.9	74.9	61.7	49.5	79.7	70.5	53.6	82.2	67.9
	MDD [25]	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	60.2	82.3	68.1
	GSP [19]	56.8	75.5	78.9	61.3	69.4	74.9	61.3	52.6	79.9	73.3	54.2	83.2	68.4
	MCD [42]	48.9	68.3	74.6	61.3	67.6	68.8	57	47.1	75.1	69.1	52.2	79.6	67.8
	DANN [1]	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
	CDAN [33]	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
	BSP [4]	52.0	68.6	76.1	58.0	70.3	70.2	58.6	50.2	77.6	72.2	59.3	81.9	66.3
	TSA [28]	55.8	73.7	79.0	61.9	74.6	74.5	60.7	53.2	80.1	72.7	58.4	84.3	69.1
	GVB-GD [9]	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
	SRDC [45]	52.3	76.3	81.0	69.5	76.2	78.0	68.7	53.8	81.7	76.3	57.1	85.0	71.3
OT-based	JDOT [5]	44.7	63.4	68.6	55.1	64.4	60.7	54.3	48.4	72.8	67.4	57.2	69.4	60.5
	DeepJDOT [11]	39.7	50.4	62.5	39.5	54.4	53.2	36.7	39.2	63.6	52.3	45.4	70.5	50.6
	MLOT [23]	52.4	72.6	77.8	58.8	72.4	72.8	53.2	50.8	76.5	70.9	54.4	82.3	66.2
	DANN [1] + MMI [28]	55.7	74.8	80.4	63.6	74.1	76.6	64.2	54.7	80.5	74.3	58.2	81.3	69.9
	DANN + MMI + COT (Ours)	57.6	75.2	83.2	67.8	76.2	75.7	65.4	56.2	82.4	75.1	60.7	84.7	71.7
	Ours + CDTrans [52]	**70.1**	**86.3**	**86.4**	**82.2**	**88.6**	**89.1**	**77.2**	**65.7**	**89.2**	**83.2**	**67.4**	**89.8**	**81.3**

methods in the Optimal Transport realm. Finally, we conduct ablative experiments together with visualization results to further demonstrate the effectiveness of our COT.

5.1. Datasets and Implementation Details

Office-31 is a benchmark dataset on the real-world unsupervised domain adaptation. It has 4110 images for 31 classes drawn from three domains: Amazon (A), DSLR (D), and Webcam (W). The 31 classes in the dataset consist of objects that commonly appeared in office settings, such as keyboards, file cabinets, and laptops.

Office-Home is a challenging benchmark dataset for domain adaptation which has 4 domains where each domain consists of 65 categories. The four domains are Art – artistic images in the form of sketches, paintings, ornamentation, etc.; Clipart – a collection of clipart images; Product – images of objects without a background and Real-World – images of objects captured with a regular camera. It contains 15,500 images in 65 classes.

VisDa-2017 is a large-scale simulation-to-real dataset for domain adaptation, which has over 280,000 images across 12 categories in the training, validation, and testing domains. The training images are generated from the same object under different circumstances, while the validation images are collected from MSCOCO ([31]).

Implementation Details Follow GVB ([10]), we adopt ResNet-50 pretrained on the ImageNet ([12]) as our backbone for Office-31 and Office-Home benchmarks and ResNet-101 for VisDa-2017 dataset. Note that our COT is a plug-and-play module, indicating it can integrate with a vast body of existing UDA approaches, e.g. DANN [1], CDAN [33]. If there is no extra statement, our COT is implemented on DANN with Maximization Mutual Information Loss (MMI) [28]. In this paper, all experiments are implemented by PyTorch. For the optimizer schedule, we adopt SGD with a momentum of 0.9. The total number of training epochs is 40,40,30 on Office-31, Office-Home, and VisDA-2017.

5.2. Results

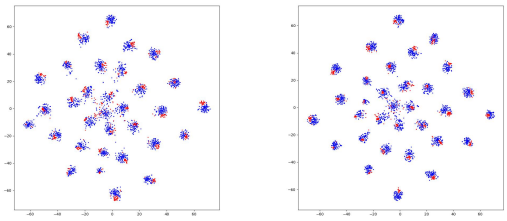
Results on Office-Home. Table 2 presents the results of OT-based and common UDA methods on Office-Home dataset. Our COT achieves the highest accuracy 71.7%, outperforming other OT-based UDA algorithms by 1.8% at least. Such tremendous improvements demonstrate that our COT can capture a more accurate local-level alignment than previous DA-based methods'. Besides, when compared with other state-of-the-art common UDA methods, to the best of our knowledge, our COT is the first OT-based UDA work that can perform a significant competitive result on the Office-Home dataset.

Results on Office-31. The results are reported in the Table 1. Our COT achieves the best performance (91.0%) both on the common and OT-based UDA.

Results on VisDA-2017. Table 3 shows the performance of OT-based and common UDA methods on the challenging VisDa-2017 dataset. Our COT outperforms the existing OT-based / common UDA methods (CNN backbone) by 3.1% / 3.2% at least, illustrating the dominant role of our COT on UDA task. Furthermore, due to the plug-and-play property of our COT, we directly add this module into the sota UDA method (CDTrans [52] on the challenging dataset VisDa2017 and Office-Home. The results in Table 3 show that our method achieves a new state-of-the-art performance, indicating the superiority and appealing potential of our COT.

Table 356 Accuracy (%) on VisDA-2017 for UDA (ResNet-101). The best result is in bold.

	Method	plane	bcybl	bus	car	horse	knife	mcyle	persn	plant	sktb	train	truck	mean
Common UDA	DANN [1]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
	MinEnt [18]	87.4	55.0	75.3	63.8	87.4	43.6	89.3	72.5	82.9	78.6	85.6	27.4	70.7
	TSA [28]	93.0	77.8	82.2	50.8	89.9	28.0	77.1	70.0	85.2	80.0	86.1	43.0	71.9
	BSP [4]	92.2	72.5	83.8	47.5	87.0	54.0	86.8	72.4	80.6	66.9	84.5	37.1	72.1
	MCC [21]	90.4	79.8	72.3	55.1	90.5	86.8	86.6	80.0	94.2	76.9	90.0	49.6	79.4
	MODEL [27]	94.8	73.4	68.8	74.8	93.1	95.4	88.6	84.7	89.1	84.7	83.5	48.1	81.6
	STAR [35]	95.0	84.0	84.6	73.0	91.6	91.8	85.9	78.4	94.4	84.7	87.0	42.2	82.7
	BNM [8]	89.6	61.5	76.9	55.0	89.3	69.1	81.3	65.5	90.0	47.3	89.1	30.1	70.4
	MSTN+DSBN [3]	94.7	86.7	76.0	72.0	95.2	75.1	87.9	81.3	91.1	68.9	88.3	45.5	80.2
	CGDM [14]	92.8	85.1	76.3	64.5	91.0	93.2	81.3	79.3	92.4	83.0	85.6	44.8	80.8
	SHOT [29]	94.3	88.5	80.1	57.3	93.1	93.1	80.7	80.3	91.5	89.1	86.3	58.2	82.9
	TVT [55]	92.9	85.6	77.5	60.5	93.6	98.2	89.4	76.4	93.6	92.0	91.7	55.7	83.9
	CDTrans [52]	**98.0**	86.9	**87.9**	80.9	97.9	**97.3**	**96.8**	85.3	97.6	83.2	**94.0**	54.4	88.4
OT-based	JDOT [5]	78.4	70.8	79.4	68.8	82.3	80.5	84.2	70.7	88.4	68.8	78.4	45.7	74.7
	DeepJDOT [11]	85.4	73.4	77.3	87.3	84.1	64.7	91.5	79.3	91.9	44.4	88.5	61.8	77.4
	MLOT [23]	88.2	70.4	77.3	50.2	84.8	77.2	80.4	74.4	83.8	68.2	82.3	38.7	73.0
	RWOT [51]	95.1	80.3	83.7	**90.0**	92.4	68.0	92.5	82.2	87.9	78.4	90.4	**68.2**	84.0
	DANN+MMI+COT (Ours)	96.9	**89.6**	84.2	74.1	96.4	96.5	88.6	82.0	96.0	**94.1**	85.1	62.1	87.1
	Ours + CDTrans [52]	**98.2**	89.4	87.6	82.3	**98.0**	97.2	96.4	**86.2**	**98.3**	92.6	92.2	58.1	**89.7**



(a) RWOT (b) COT

Figure 1. t-SNE of classifier responses by RWOT and COT on CI50 (red: Amazon, blue: Webcam).

5.3. Ablation study

Ability on Resisting Class-imbalance Challenge. As the results illustrated above, our COT brings little gain than the existing sota OT-based UDA method (RWOT [51]) on the Office-31 dataset while achieving the dominant performance on Office-Home and VisDA-2017 datasets. To find the reason for this, we compute the label distribution of the Office-31 dataset, where the class imbalance phenomenon is not occurred obviously compared to the Office-Home and VisDA-2017 datasets (shown in Fig. 2). Thus, to verify the ability of our COT on resisting class-imbalance challenges among all UDA benchmark datasets, we construct 3 types of class-imbalance evaluation benchmarks by randomly erasing 30%, 50%, 70% samples from certain classes on Office-31 datasets target domain, termed CI30, CI50, and CI70, respectively. Table 4 presents the results of our COT and existing OT-based UDA methods on Office-31, CI30, CI50, and CI70. According to this quantitative result, we discover that the more severe the class imbalance phenomenon occurs, the larger the performance gap between our COT and RWOT, demonstrating that our COT embraces excellent ability on handling class-imbalance challenges.

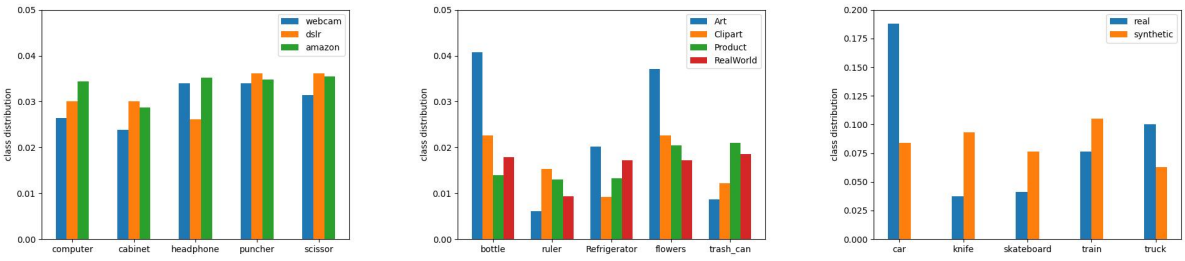
Table 4. Mean accuracy of OT-based UDA methods on class-imbalance datasets.

Dataset	COT	RWOT	MLOT
Office-31	**91.0**	90.8	87.8
CI30	**88.2**	84.7	83.6
CI50	**87.7**	78.3	80.4
CI70	**85.4**	70.5	75.2

Computation Cost. Here, we compute the computation cost of our COT and RWOT, where RWOT is a representative work of a sample-level OT matching mechanism. By looking at the results in Fig. 3, compared with RWOT, our COT significantly saves the computational cost with the increasing of training scale. Integrating the above theoretical analysis with this quantitative result demonstrates the superiority of our COT in economizing the computation head.

Effects of Loss Weight. α_1 and α_2 are employed to weight the importance of cluster and OT loss, respectively. In our experiment, α_1 and α_2 are the same and selected from a broad range of $\{0.1, 0.2, 0.4, 1.0, 2.0\}$. Results are evaluated on the Office-31 and shown in Table 5. When α_1 and α_2 are set to large value, the performance will tend to drop. In our opinion, the larger α_1 and α_2 would put more attention to local alignment while weakening the feature representation on both domains. As a result, α_1 and α_2 are set to 0.2, 0.2 for all datasets.

Effects of the Number of Sub-domains. Y, defined in Section 4, represents the number of sub-domains for each



(a) Office-31 (b) Office-Home (c) VisDA-2017

Figure 2. Class distribution on Office-31, Office-Home and VisDA-2017. Note that the visualized class distribution is randomly selected from corresponding datasets.

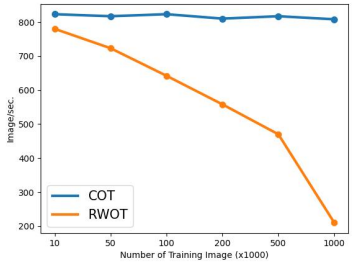


Figure 3. Computation cost of COT and RWOT

Table 5. Mean accuracy of our COT with different α_1 and α_2 on Office-31 dataset.

α_1	α_2	avg_acc
0.1	0.1	90.8
0.2	**0.2**	**91.0**
0.4	0.4	90.6
1.0	1.0	89.4
2.0	2.0	87.7

class. We investigate a broad range of Y to find an appropriate value. Based on the results shown in Table 6, we set Y to 2 for all test settings.

Table 6. Mean accuracy of our COT with different Y on Office-31 dataset.

Y	avg_acc
1	90.7
2	91.0
3	90.8
5	90.4

Qualitative Analysis on COT As discussed above, we present 2 advantages of our COT when compared with existing OT-based methods [23, 51]. In this part, we further illustrate another advantage compared to the clustering-based OT algorithm in the conventional OT realm [24, 36]. Our COT is optimized in an end-to-end manner, thus the feature representation and optimal transportation map are both updated asynchronously. While for [24] and [36], they focus on obtaining offline cluster centers firstly and then compute the optimal transportation map. Such a synchronous optimization manner inevitably leads to a sub-optimal mapping result, because the computation procedure of an optimal transportation map is based on the fixed cluster centers. This is why we adopt an online clustering mode instead of an offline one.

5.4. Visualization

We present the t-SNE visualization of feature representation for our COT and RWOT [51] on CI50. Under the challenge of class imbalance, the learned features from RWOT in some classes are ambiguous to provide accurate representation compared to our COT. These visualization results further demonstrate the excellent ability of our COT on handling class-imbalance challenges from the feature representation perspective.

6. Conclusion

In this paper, we propose a novel method to advance the OT-based domain adaptation community, which integrates optimal transport theory with clustering operation, termed Clustering-based Optimal Transport (COT). Concretely, COT applies the loss derived from discrete Kantorovich dual form to cluster and align centers in the source and target domain, thus transferring knowledge from the source domain to the target domain. Moreover, our COT can eliminate the negative effect brought by the class imbalance phenomenon and reduce the computation cost simultaneously, which are two challenging problems for existing OT-based UDA algorithms. Besides, we also provide comprehensive theoretical analysis and experimental results to guarantee the advantages of our COT.

References

[1135 Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014. 1, 5, 6, 7

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. 2017. 3

[3] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *CVPR*, 2019. 7

[4] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning*, pages 1081–1090. PMLR, 2019. 6, 7

[5] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *arXiv preprint arXiv:1705.08848*, 2017. 2, 5, 6, 7

[6] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014. 1

[7] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016. 2

[8] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3941–3950, 2020. 6, 7

[9] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Tian Qi. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 5, 6

[10] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12455–12464, 2020. 6

[11] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018. 5, 6, 7

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1, 3

[14] Zhekai Du, Jingjing Li, Hongzu Su, Lei Zhu, and Ke Lu. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. 2021. 7

[15] Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR, 2021. 1

[16] Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996. 3

[17] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2020. 2

[18] Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. In *CAP*, pages 281–296, 2005. 7

[19] Sahand Hajifar and Hongyue Sun. Online domain adaptation for continuous cross-subject liver viability evaluation based on irregular thermal data, 2020. 5, 6

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016. 1

[21] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation, 2020. 5, 7

[22] Leonid V Kantorovich. On the translocation of masses. *Journal of mathematical sciences*, 133(4):1381–1382, 2006. 2

[23] Tanguy Kerdoncuff, Rémi Emonet, and Marc Sebban. Metric learning in optimal transport for domain adaptation. 2020. 1, 2, 5, 6, 7, 8

[24] John Lee, Max Dabagia, Eva Dyer, and Christopher Rozell. Hierarchical optimal transport for multimodal distribution alignment. *Advances in Neural Information Processing Systems*, 32, 2019. 8

[25] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2020. 5, 6

[26] Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. Enhanced transport distance for unsupervised domain adaptation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13933–13941, 2020. 6

[27] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020. 7

[28] Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. *arXiv preprint arXiv:2103.12562*, 2021. 1, 5, 6, 7

[29] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. 2020. 5, 7

[30] Chi-Heng Lin, Mehdi Azabou, and Eva L Dyer. Making transport more robust and interpretable by moving data

through a small number of anchor points. *Proceedings of machine learning research*, 139:6631, 2021. 1

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[32] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pages 4013–4022. PMLR, 2019. 5, 6

[33] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017. 5, 6

[34] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 5, 6

[35] Zhihe Lu, Yongxin Yang, Xiatian Zhu, Cong Liu, Yi-Zhe Song, and Tao Xiang. Stochastic classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9111–9120, 2020. 7

[36] Liang Mi, Wen Zhang, and Yalin Wang. Regularized wasserstein means for aligning distributional data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5166–5173, 2020. 8

[37] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781. 2

[38] Khai Nguyen, Dang Nguyen, Quoc Nguyen, Tung Pham, Hung Bui, Dinh Phung, Trung Le, and Nhat Ho. On transportation of mini-batches: A hierarchical approach. *arXiv preprint arXiv:2102.05912*, 2021. 1

[39] Yingwei Pan, Ting Yao, Yehao Li, Yu Wang, Chong-Wah Ngo, and Tao Mei. Transferrable prototypical networks for unsupervised domain adaptation, 2019. 6

[40] Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 849–858. PMLR, 2019. 1, 2

[41] Hochang Rhee and Nam Ik Cho. Efficient and robust pseudo-labeling for unsupervised domain adaptation. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 980–985, 2019. 2

[42] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 1, 5, 6

[43] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018. 5

[44] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015. 1

[45] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8725–8735, 2020. 5, 6

[46] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 5

[47] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 3

[48] Qian Wang and Toby Breckon. Unsupervised domain adaptation via structured prediction based selective pseudo-labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6243–6250, 2020. 2

[49] Wei Wang, Haojie Li, Zhengming Ding, and Zhihui Wang. Rethink maximum mean discrepancy for domain adaptation. *arXiv preprint arXiv:2007.00689*, 2020. 1

[50] Pengfei Wei, Yiping Ke, Xinghua Qu, and Tze-Yun Leong. Subdomain adaptation with manifolds discrepancy alignment, 2020. 1

[51] Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4394–4403, 2020. 1, 5, 7, 8

[52] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021. 6, 7

[53] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2272–2281, 2017. 1

[54] Yuguang Yan, Wen Li, Hanrui Wu, Huaqing Min, Mingkui Tan, and Qingyao Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *IJCAI*, volume 7, pages 2969–2975, 2018. 1

[55] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2108.05988*, 2021. 7

[56] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation, 2019. 6

[57] Yongchun Zhu, Fuzhen Zhuang, Jindong Wang, Guolin Ke, Jingwu Chen, Jiang Bian, Hui Xiong, and Qing He. Deep subdomain adaptation network for image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1713–1722, Apr 2021. 1