# UNSUPERVISED DOMAIN ADAPTATION BY OPTIMAL TRANSPORTATION OF CLUSTERS BETWEEN DOMAINS

**Anonymous authors**
Paper under double-blind review

## 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 borrow the experience from Optimal Transport (OT) theory to construct instance-pair alignment. However, existing OT-based algorithms are limited to resolve class imbalance challenge and require a huge computation cost when considering a large-scale training situation. In this paper, we address these two issues by proposing a Clustering-based Optimal Transport (COT) algorithm, which formulates the alignment procedure as an Optimal Transport problem by capturing the fine-grained attribute alignment. Concretely, COT innovatively designs the loss derived from discrete Kantorovich dual form to construct a mapping between clustering centers in source and target domain, which simultaneously eliminates the negative effect brought by class imbalance and reduces the computation cost on the basis of theoretical analysis. Finally, our COT together with some previous UDA methods achieve superior performance on several benchmarks.

## 1 INTRODUCTION

Benefiting from the availability of large-scale data , the field of 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 shift challenge, resulting in the spurious predictions on the new domain. Furthermore, collecting labeled data in various domains is labor-intensive and expensive. To alleviate the negative effect brought by the domain discrepancy, Unsupervised Domain Adaptation (UDA) has attracted many researchers' attention, which can transfer the knowledge from a labeled domain to an unlabeled domain.

Previous unsupervised domain adaptation methods (Yan et al. (2017); Saito et al. (2018); Wang et al. (2020)) mainly seek to learn a global domain shift by aligning the global source and target distributions, while ignoring the local-level alignment between two domains. Under the guidance of global domain adaptation, the distributions of source and target domain are almost the same, thus losing the fine-grained information for each class in target domain. This would be a fatal problem in the existing global domain adaptation methods.

In order to bridge the local alignment between source and target domain, recent researchers employs the experience from Optimal Transport theory to construct instance-pair alignment between domains. Compared with traditional global alignment algorithms, OT-based UDA methods can preserve the domain-specific properties since the instance-level alignment is highlighted. However, there exist two drawbacks on recent OT-based UDA algorithms. 1) When considering a realistic situation, i.e. the class imbalance[1] phenomenon occurs between source and target domain, samples belong to the same class in the target domain are assigned with different pseudo labels due to the mechanism of optimal transport, which requires each sample in source domain has to be mapped on target samples under the constraint of marginal distribution preservation. As a result, current OT-based UDA methods fail to provide accurate local alignment between source and target domain

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

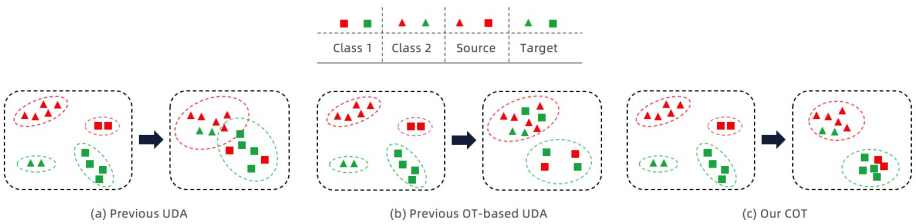



Figure 1: Comparison of previous UDA methods and our proposed method when facing class-imbalance challenge. (a) Previous UDA methods aim to align the source and target domain from the global perspective, leading the ambiguous boundary between different classes. (b) Traditional OT-based UDA algorithms further align the feature representation on the instance-level, which performs poorly in the face of class-imbalance challenge due to the false-assigned label on the sample within the target domain. (c) Our COT presents a cluster-based optimal transport alignment algorithm, which can get accurate fine-grained representation to resist the class-imbalance risk.

when facing class imbalance challenge. 2) Previous OT-based UDA methods aim to find a sample-level optimal counterpart, which wastes a large amount of computation cost, especially training on the large-scale benchmarks.

To resolve two aforementioned drawbacks on the OT-based UDA algorithms, we propose a Clustering-based Optimal Transport (COT) algorithm to construct a mapping between clustering centers in source and target domain. Instead of aligning the feature representation between domains on the instance level, COT align the clustering centers from source to target domain by applying optimal transport metric. Clusters in source domain are obtained from the classifiers supervised by the labeled source domain data. For target domain, COT utilizes a set of learnable clusters to represent the feature distribution of target domain, which can describe the sub-domain information (Zhu et al. (2021); Wei et al. (2020)) implicitly. The clusters on source and target domain can represent the individual sub-domains information respectively, such that optimal transport between clusters intrinsically provide a local mapping from sub-domains in source domain to those in target domain. To this end, our Cluster-based Optimal Transport method, achieving the distinct local alignment and capturing amazing abilities to resist class-imbalance challenge (shown in Fig. 1) and reduce computation cost, facilitates the OT-based UDA community significantly. Moreover, we provide a theoretical analysis to guarantee that 1) COT can eliminate the negative effect brought by class imbalance 2) COT mitigates much computation cost than previous OT-based UDA methods.

In summary, our main contributions include: 1) We propose a novel Cluster-based Optimal Transport module with a special designed loss derived from discrete type of Kantorovich dual form, which aligns the clusters between source and target domain to transfer knowledge between individual sub-domains (clusters) and resist the class imbalance challenge; 2) Our COT can be more efficient on the practical application in the field of UDA by aligning each cluster instead of individual sample, which economizes the computation cost effectively; 3) Our COT can be well integrated with existing unsupervised domain adaptation methods. Empirically, COT together with MCC (Jin et al. (2020)), DANN (Ajakan et al. (2014)), CDAN (Long et al. (2017a)) achieve state-of-the-art performance on several benchmarks.

## 2 Related Work

**Pseudo Label based Domain Adaptation** Supervised domain adaptation methods utilize the class label information as a guide for efficiently transferring knowledge between different domains, which assumes that fine-tuning the deep neural network model with labeled source data can remedy the domain shift. When labeled target data are unavailable for unsupervised domain adaptation task, several methods have been proposed to substitute for labeled data by introducing pseudo labels.

Inspired by the observation that samples in target domain can be clustered within the feature space, for accurate pseudo-labeling, Wang & Breckon (2020) propose a selective pseudo-labeling strategy based on structural predictions which utilize the unsupervised clustering analysis. Rhee &

Cho (2019) 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 can enable the performance less sensitive to threshold which determine the pseudo-labels. In the task of person re-identification, Ge et al. (2020) 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. Morerio et al. (2020) provide an characterization of shift noise and show that the conditional Generative Adversarial Networks (cGANs) are robust to shift noise to some extent. Specifically, the generator allows for cleaner samples from target distribution and classifier allows for better label assignment for target samples.

**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, Monge (1781) put forward the optimal transport problem for the first time. Kantorovich (2006) provide an extension of the original problem of Monge. Recently, new computation strategy have emerged and make possible the application of optimal transport in domain adaptation.

Courty et al. (2016) 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 same class must undergo similar transformation. Courty et al. (2017) minimize the optimal transport loss between the joint source distribution and the estimated target joint distribution depending on a function which is introduced to predict an output value given input from source domain. For reducing discrepancy between multiple domains, Redko et al. (2019) 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 Kerdoncuff et al. (2020) which optimize the metric for domain adaptation. Taking the ignorance of intra-domain structure of current domain adaptation based on optimal transport, Xu et al. (2020) focus on the target samples distributed near the edge of clusters/far from corresponding class centers which may be easily misclassified by the decision boundary learned from source domain. The proposition of Shrinking Subspace Reliability exploits spatial prototypical information and intra-domain structure to dynamically measure the sample-level domain discrepancy across domains.

## 3 PRELIMINARY

In this section, we will 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 $\{y_i^s\}_{i=1}^{n_s}$ with $y_i^s \in \mathcal{Y}$. 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 which 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*

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

*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*

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

*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:*

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

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

$$W_c(\mu_{\mathcal{S}}, \mu_{\mathcal{T}}) = \max_{\varphi,\psi}\{\int_{X_{\mathcal{S}}} \varphi(x^s)d\mu_{\mathcal{S}}(x^s) + \int_{X_{\mathcal{T}}} \psi(x^t)d\mu_{\mathcal{T}}(x^t)|\varphi(x^s) + \psi(x^t) \leq c(x^s, x^t)\} \tag{4}$$

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

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

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 $\varphi$.

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

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

where $\varphi$ is under the constraint that $|\varphi(x) - \varphi(x')| \leq \|x - x'\|_2$. WGAN Arjovsky et al. (2017) 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 $\varphi$ is bounded from above by 1. When we set the cost function as $c(x^s, x^t) = \|x^s - x^t\|_2^2$, by Gangbo & McCann (1996) that the existence and uniqueness of optimal transport map is guaranteed.

# 4 METHOD

## 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 source and target domain by ImageNet pretrained CNNs. Then we utilize learnable clusters to represent the sub-domains in 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 pretrained (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 feature vary during the training due to the parameters updating in the feature extractor.

**Clustering** As for each sample from source domain, i.e., $x_i^s$ has a label $y_i^s \in \mathcal{Y}$. 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:

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

For source domain, we take the classifiers $\{w_v^s\}_{v=1}^{|\mathcal{Y}|}$ as clusters for feature space of source domain. For 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 hyper-parameter which stands for the number of sub-domains in target domain, $Q$ is a positive integer which represents the number of sub-domains for each class.

**Algorithm 1** Clustering based on Optimal Transport

Set number of epochs for training as $E$, updating duration for COT as $f_{\text{ot}}$, learnable clusters for target domain as $\{w_u^t\}_{u=1}^{|\mathcal{Y}|}$, classifiers/clusters for source domain as $W = \{w_v^s\}_{v=1}^{|\mathcal{Y}|}$;
**for** k-th training epoch while $k \leq E$ **do**
  **for** $l$-th iteration in $k$-th epoch **do**
    1. Take mini-batch of samples from source and target domain as input for feature extractor CNNs with parameters $\theta$, the output features are $\{x_i^s\}_{i=1}^b$ and $\{x_j^t\}_{j=1}^b$;
    2. compute the $L_{\text{cluster}}$ for $\{x_j^t\}_{j=1}^b$ and $L_{\text{cross-entropy}}$ for $\{x_i^s\}_{i=1}^b$ in the $l$-th batch;
    3. compute the clustering based optimal transport loss $L_{\text{OT}}$;
    **if** $1 \leq (v \mod 10) \leq f_{\text{ot}}$ **then**
      we find the current optimal map $s$ from source domain to the set of clusters in target domain by maximizing $L_{\text{OT}}$ and update $\{\lambda_v^t\}_{v=1}^{|\mathcal{Y}|}$;
    **end if**
    4. optimize the following weighted loss

$$\alpha_1 \cdot L_{\text{cross-entropy}} + \alpha_2 \cdot L_{\text{cluster}} + \alpha_3 \cdot L_{\text{OT}}$$

    update $\{w_u^t\}_{u=1}^{|\mathcal{Y}|}$,$\{w_v^s\}_{v=1}^{|\mathcal{Y}|}$ and parameters $\theta$, where $\alpha_1, \alpha_2, \alpha_3$ are weights.
  **end for**
**end for**

$\{w_u^t \in \mathbb{R}^c\}_{u=(l-1)\cdot Q+1}^{l \cdot Q}$ 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 suitable metric to measure the distance between features and clusters and pull the features back to corresponding clusters.

$$L_{\text{cluster}} = \text{distance}(\{w_u^t\}_{u=1}^K, \{x_j^t\}_{j=1}^{n_t}) \tag{8}$$

**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:

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

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).

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

where $\psi^c(w_v^s) = \inf_{u=1}^{K} (c_{vu} - \psi(w_u^t))$. We seek for the optimal transportation map between clusters by optimizing the following loss.

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

where $\{\lambda_u^t\}_{u=1}^K$ represent the value of function $\psi$ at points $\{w_u^t\}_{u=1}^K$. Furthermore, it is worth noting that cost $c_{vu}$ is frozen during the optimization of $L_{\text{OT}}$. Whenever we update the parameters of feature extractor CNNs, $c_{vu}$ will be updated in the meantime. The details of our COT's optimization strategy of is shown in Algorithm 1.

### 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

$$\begin{aligned} &\min_{T\in\mathbb{R}^{n_s\times n_t}} T_{ij}c_{ij} \\ &s.t.\sum_{j=1}^{n_t} T_{ij} = \frac{1}{n_s}, \sum_{i=1}^{n_s} T_{ij} = \frac{1}{n_t}, T_{ij} \geq 0, \forall 1 \leq i \leq n_s, 1 \leq j \leq n_t. \end{aligned} \tag{12}$$

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

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

$$\begin{aligned} &\|x_i^s - w_{v_1^s}\|_2^2 - \|x_i^s - w_{v_2}^s\|_2^2 \\ =&\|w_{v_1}^s\|_2^2 - \|w_{v_2}^s\|_2^2 + 2(\|w_{v_2}^s\|_2 x_i^{s\top}\frac{w_{v_2}^s}{\|w_{v_2}^s\|_2} - \|w_{v_1}^s\|_2 x_i^{s\top}\frac{w_{v_1}^s}{\|w_{v_1}^s\|_2}) \end{aligned} \tag{13}$$

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 draw the conclusion that the similarity between features and clusters are almost equivalent to distance between features and classifiers. With labels as supervision, the optimization of cross-entropy can promote the inter-class discrepancy which imply

$$x_i^{s\top}\frac{w_{y_i^s}^s}{\|w_{y_i^s}^s\|_2} \gg x_i^{s\top}\frac{w_v^s}{\|w_v^s\|_2}, \forall v \neq y_i^s \tag{14}$$

which also provide the following result

$$\|x_i^s - w_{y_i^s}^s\|_2^2 \ll \|x_i^s - w_v^s\|_2^2, \forall v \neq y_i^s \tag{15}$$

If clustering doesn't work sufficiently well, it happens that some samples in source domain with label $v$ are assigned to samples in target domain with label $u \neq v$. When clustering perform 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

$$\sum_{i,j} T_{ij}c_{ij} \sim \sum_{v,j}(\sum_{x_i^s\in X_v^s} T_{ij})\bar{c}_{vj} \tag{16}$$

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

$$\sum_{j=1}^{n_t}(\sum_{x_i^s\in X_v^s} T_{ij}) = \frac{n_v}{n_s}, \sum_{v=1}^{|\mathcal{Y}|}(\sum_{x_i^s\in X_v^s} T_{ij}) = \frac{1}{n_t} \tag{17}$$

Inspired by Equation (16) and (17), we consider the following optimal transport between clusters from source domain and instances from target domain instead of solving Kantorovich problem in Equation (12).

$$\begin{aligned} &\min_{\bar{T}\in\mathbb{R}^{|\mathcal{Y}|\times n_t}} \bar{T}_{vj}\bar{c}_{vj} \\ &s.t.\sum_{j=1}^{n_t} \bar{T}_{vj} = \frac{n_v}{n_s}, \sum_{v=1}^{|\mathcal{Y}|} \bar{T}_{vj} = \frac{1}{n_t}, \bar{T}_{vj} \geq 0, \forall 1 \leq v \leq |\mathcal{Y}|, 1 \leq j \leq n_t \end{aligned} \tag{18}$$

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

$$\exists\delta > 0, \text{ s.t. } \|(\frac{n_1^s}{n_s}, ..., \frac{n_{|\mathcal{Y}|}^s}{n_s}) - (\frac{n_1^t}{n_t}, ..., \frac{n_{|\mathcal{Y}|}^t}{n_t})\|_2 \geq \delta \tag{19}$$

where $\delta$ is a constant which measure 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 source domain will be assigned to samples in target domain with label $u \neq v$. This will result in samples

belonging to same category in the target domain be given different pseudo labels, which increase the difficulty of training and cause the degradation of performance of deep learning methods on target domain.

When we utilize the clustering based optimal transport, for source domain, we have $\sum_{u=1}^{K} T_{vu} = \frac{1}{|\mathcal{Y}|}$. For target domain, $\sum_{v=1}^{|\mathcal{Y}|} \sum_{u=(l-1)\cdot Q+1}^{l\cdot Q} T_{vu} = \frac{Q}{K} = \frac{1}{|\mathcal{Y}|}$, 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 source and target domain, computation cost on feature extractor is shown as follows:

$$\mathcal{O}(n_s + n_t) \cdot \mathcal{O}(\text{feature-extractor}) \tag{20}$$

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

$$\mathcal{O}(n_s \cdot n_t) \tag{21}$$

In comparison, for cluster-based optimal transport, the main computation cost is on optimal transport:

$$\mathcal{O}(|\mathcal{Y}| \cdot K) \tag{22}$$

For a large scale dataset, clustering-based optimal transport cost much less than instance-based optimal transport.

## 5 Experiments

In this section, we compare our method with state-of-the-art unsupervised domain adaptation methods on the three authoritative benchmarks,including Office-31, Office-Home and VisDa-2017.

### 5.1 Datasets and implementation detail

**Office-31** is a famous 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 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 – 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 (Lin et al. (2014)).

**Implementation details** Followed by GVB (Cui et al. (2020c)), we adopt ResNet-50 pretrained on the ImageNet (Deng et al. (2009)) as our backbone for Office-31 and Office-Home benchmarks and ResNet-101 for VisDa-2017 dataset. In this paper, all experiments are implemented by PyTorch. For optimizer schedule, we adopt SGD with momentum with 0.9. We apply our method on the DANN and MCC respectively. For each transferring task, we report the average accuracy of 3 random trails. The optimal transport loss weight, cluster loss weight and classification loss weight are $0.1, 0.1$ and 1 respectively.

### 5.2 RESULTS

**Results on Office-31.** The results are shown in the Table 1. We implement our method on the basis of DANN and MCC. When comparing with other state-of-the-art methods, MCC + Ours achieves the highest average accuracy 90.7%. Besides, DANN + Ours brings a remarkable enhancement of 8.2% to DANN, demonstrating that our method is complementary with adversarial-based domain adaptation method. Our COT achieves the most significant enhancement on the Office-31 benchmark, which embraces the most severe class-imbalance situation comparing with other evaluated datasets.

Table 1: Accuracy (%) on Office-31 for UDA (ResNet-50).

| Method | A→W | D→W | W→D | A→D | D→A | W→A | Avg |
|---|---|---|---|---|---|---|---|
| ADDA Tzeng et al. (2017) | 86.2 | 96.2 | 98.4 | 77.8 | 69.5 | 68.9 | 82.9 |
| JAN Long et al. (2017b) | 85.4 | 97.4 | 99.8 | 84.7 | 68.6 | 70.0 | 84.3 |
| MCD Saito et al. (2018) | 88.6 | 98.5 | **100.0** | 92.2 | 69.5 | 69.7 | 86.5 |
| BNM Cui et al. (2020a) | 91.5 | 98.5 | **100.0** | 90.3 | 70.9 | 71.6 | 87.1 |
| AFN Xu et al. (2019) | 88.8 | 98.4 | 99.8 | 87.7 | 69.8 | 68.7 | 85.7 |
| DMRL Wu et al. (2020) | 90.8 | 99.0 | **100.0** | 93.4 | 73.0 | 71.2 | 87.9 |
| GTA Sankaranarayanan et al. (2018) | 89.5 | 97.9 | 99.8 | 87.7 | 72.8 | 71.4 | 86.5 |
| SymNets Zhang et al. (2019) | 90.8 | 98.8 | **100.0** | 93.9 | 74.6 | 72.5 | 88.4 |
| CDAN Long et al. (2017a) | 94.1 | 98.6 | **100.0** | 92.9 | 71.0 | 69.3 | 87.7 |
| TAT Liu et al. (2019) | 92.5 | 99.3 | **100.0** | 93.2 | 73.1 | 72.1 | 88.4 |
| MDD Li et al. (2020a) | 94.5 | 98.4 | **100.0** | 93.5 | 74.6 | 72.2 | 88.9 |
| GVB-GD Cui et al. (2020b) | 94.8 | 98.7 | **100.0** | 95.0 | 73.4 | 73.7 | 89.3 |
| GSP Hajifar & Sun (2020) | 92.9 | 98.7 | 99.8 | 94.5 | 75.9 | 74.9 | 89.5 |
| TSA Li et al. (2021) | 96.0 | 98.7 | **100.0** | 95.4 | 76.7 | 76.8 | 90.6 |
| DANN Ajakan et al. (2014) | 82.0 | 96.9 | 99.1 | 79.7 | 68.2 | 67.4 | 82.2 |
| **+Ours** | 95.2 | 98.6 | **100.0** | 94.4 | **76.7** | **77.4** | 90.4 |
| MCC Jin et al. (2020) | 95.5 | 98.6 | **100.0** | 94.4 | 72.9 | 74.9 | 89.4 |
| **+Ours** | **96.5** | **99.1** | **100.0** | **96.1** | 76.5 | 76.1 | **90.7** |

**Results on Office-home.** The results are reported in the Table 2. We introduce our method on the basis of DANN and achieves the highest average accuracy 70.6%. Note that the result of TSA (Li et al. (2021)) is re-implemented by the official code.

Table 2: Accuracy (%) on Office-Home for UDA (ResNet-50).

| 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| JAN Long et al. (2017b) | 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 Liu et al. (2019) | 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 Pan et al. (2019) | 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 Li et al. (2020b) | 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 Zhang et al. (2019) | 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 Cui et al. (2020a) | 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 Li et al. (2020a) | 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 Hajifar & Sun (2020) | 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 Saito et al. (2018) | 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 |
| CDAN Long et al. (2017a) | 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 Chen et al. (2019) | 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 Li et al. (2021) | 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 Cui et al. (2020b) | 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 |
| DANN Ajakan et al. (2014) | 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 |
| **+Ours** | **57.6** | **74.8** | 80.4 | 63.6 | 74.1 | **75.7** | 65.0 | 54.7 | **81.0** | **75.1** | **60.7** | **84.7** | **70.6** |

**Results on VisDA-2017.** The results are presented in the Table 3. Similar in the Office-home, we reported the result of our method on the basis of DANN. Our method outperforms MCC by 2.2% average accuracy. The result of TSA on the VisDA-2017 dataset is also re-implemented by the official code.

### 5.3 ABLATION STUDY

In this subsection, we evaluate the effectiveness of our COT module on three famous unsupervised domain adaptation algorithm, including DANN, CDAN and MCC. All experiments are conducted on the Office-31 benchmark. The significant improvement on these methods demonstrate the plug and play property of our proposed optimal transport module.

Table 3: Accuracy (%) on VisDA-2017 as *regularizer* for UDA (ResNet-101).

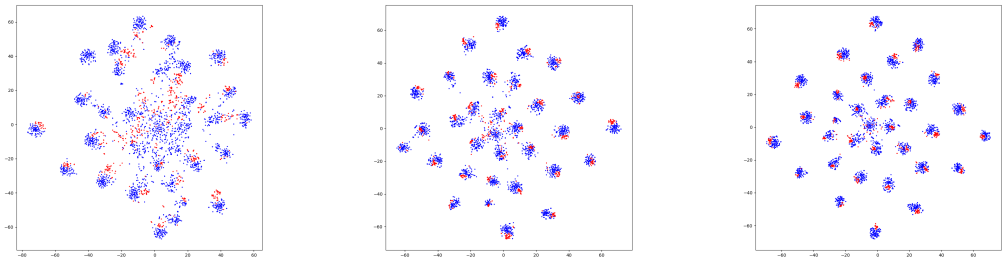
| Method | plane | bcybl | bus | car | horse | knife | mcyle | persn | plant | sktb | train | truck | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DANN Ajakan et al. (2014) | 81.9 | 77.7 | 82.8 | 44.3 | 81.2 | 29.5 | 65.1 | 28.6 | 51.9 | 54.6 | 82.8 | 7.8 | 57.4 |
| DANN + MinEnt Grandvalet et al. (2005) | 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 |
| DANN + TSA Li et al. (2021) | 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 |
| DANN + BSP Chen et al. (2019) | 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 |
| DANN + MCC Jin et al. (2020) | 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** |
| **DANN + Ours** | **93.6** | 78.1 | 82.2 | **74.1** | **91.1** | **91.8** | 88.1 | 76.8 | 88.7 | 76.6 | 84.3 | 42.3 | **80.6** |

Table 4: Ablation Study of COT Module on Office-31.

| method | A→W | D→W | W→D | A→D | D→A | W→A | Avg |
|---|---|---|---|---|---|---|---|
| DANN Ajakan et al. (2014) | 82.0 | 96.9 | 99.1 | 79.7 | 68.2 | 67.4 | 82.2 |
| **+COT** | 95.2 | 98.6 | **100.0** | 94.4 | **76.7** | **77.4** | 90.4 |
| CDAN Long et al. (2017a) | 94.1 | 98.6 | **100.0** | 92.9 | 71.0 | 69.3 | 87.7 |
| **+COT** | 94.8 | 98.8 | **100.0** | 94.4 | 74.8 | 75.4 | 89.6 |
| MCC Jin et al. (2020) | 95.5 | 98.6 | **100.0** | 94.4 | 72.9 | 74.9 | 89.4 |
| **+COT** | **96.5** | **99.1** | **100.0** | **96.1** | 76.5 | 76.1 | **90.7** |

### 5.4 VISUALIZATION

Most recent domain adaptation methods minimize the divergence between these two domains in the same embedding feature space. It turn out to be an overly rough operation which will lead to the lack of domain-specific feature representation. Instead of aligning features from different domains on instance level, we implement the alignment between clusters from source and target domains, which should be more efficient and robust. The visualization of feature cluster in source and target domain are shown in Figure 1. Obviously, our COT module together with DANN has a more compact cluster centers than other two methods.



(a) ResNet50 (b) DANN (c) COT

Figure 2: t-SNE of classifier responses by ResNet50, DANN and COT (red: Amazon, blue: Webcam).

## 6 CONCLUSION

In this paper, we propose a novel module for domain adaptation which integrates with optimal transport and clustering operation, termed as clustering based optimal transport (COT). With pseudo labels provided by learnable clusters, COT can reduce the intra-class distance and enlarge inter-class distance simultaneously. COT apply the loss derived from discrete Kantorovich dual form to cluster centers in source and target domain, thus transfering knowledge from source domain to target domain. Besides, our COT can eliminate the negative effect from class imbalance and reduce the computation cost in optimal transport. Additionally, COT is plug and play which can be well integrated with existing domain adaptation methods. Empirically, COT together with MCC, DANN, CDAN achieve state-of-the-art performance on several benchmarks, including Office-31, Office-home and VisDa-2017.

## REFERENCES

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

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. 2017.

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*, pp. 1081–1090. PMLR, 2019.

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.

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

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*, pp. 3941–3950, 2020a.

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*, 2020b.

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*, pp. 12455–12464, 2020c.

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*, pp. 248–255. Ieee, 2009.

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

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. URL https://openreview.net/forum?id=rJlnOhVYPS.

Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. In *CAP*, pp. 281–296, 2005.

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

Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation, 2020.

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

Tanguy Kerdoncuff, Rémi Emonet, and Marc Sebban. Metric learning in optimal transport for domain adaptation. 2020.

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*, pp. 1–1, 2020a. ISSN 1939-3539. doi: 10.1109/tpami.2020.2991050. URL http://dx.doi.org/10.1109/TPAMI.2020.2991050.

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)*, pp. 13933–13941, 2020b. doi: 10.1109/CVPR42600.2020.01395.

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.

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*, pp. 740–755. Springer, 2014.

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*, pp. 4013–4022. PMLR, 2019.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *arXiv preprint arXiv:1705.10667*, 2017a.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pp. 2208–2217. PMLR, 2017b.

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

Pietro Morerio, Riccardo Volpi, Ruggero Ragonesi, and Vittorio Murino. Generative pseudo-label refinement for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3130–3139, 2020.

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

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*, pp. 849–858. PMLR, 2019.

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)*, pp. 980–985, 2019. doi: 10.1109/APSIPAASC47483.2019.9023239.

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*, pp. 3723–3732, 2018.

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*, pp. 8503–8512, 2018.

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*, pp. 7167–7176, 2017.

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, pp. 6243–6250, 2020.

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

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

Yuan Wu, Diana Inkpen, and Ahmed El-Roby. Dual mixup regularized learning for adversarial domain adaptation. In *European Conference on Computer Vision*, pp. 540–555. Springer, 2020.

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*, pp. 4394–4403, 2020.

Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1426–1435, 2019.

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*, pp. 2272–2281, 2017.

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

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. ISSN 2162-2388. doi: 10.1109/tnnls.2020.2988928. URL http://dx.doi.org/10.1109/TNNLS.2020.2988928.

# A Appendix

You may include other additional sections here.