Class Overwhelms: Mutual Conditional Blended-Target Domain Adaptation

Pengcheng Xu[1], **Boyu Wang**[1, 2 *], **Charles Ling**[1]

[1] Western University, London, ON N6A 5B7, Canada
[2] Vector Institute, Toronto, ON M5G 1M1, Canada
pxu67@uwo.ca, bwang@csd.uwo.ca, charles.ling@uwo.ca

Abstract

Current methods of blended targets domain adaptation (BTDA) usually infer or consider domain label information but underemphasize hybrid categorical feature structures of targets, which yields limited performance, especially under the label distribution shift. We demonstrate that domain labels are not directly necessary for BTDA if categorical distributions of various domains are sufficiently aligned even facing the imbalance of domains and the label distribution shift of classes. However, we observe that the cluster assumption in BTDA does not comprehensively hold. The hybrid categorical feature space hinders the modeling of categorical distributions and the generation of reliable pseudo labels for categorical alignment. To address these, we propose a categorical domain discriminator guided by uncertainty to explicitly model and directly align categorical distributions $P(Z|Y)$. Simultaneously, we utilize the low-level features to augment the single source features with diverse target styles to rectify the biased classifier $P(Y|Z)$ among diverse targets. Such a mutual conditional alignment of $P(Z|Y)$ and $P(Y|Z)$ forms a mutual reinforced mechanism. Our approach outperforms the state-of-the-art in BTDA even compared with methods utilizing domain labels, especially under the label distribution shift, and in single target DA on DomainNet.

Introduction

Deep learning suffers a serious performance drop under the distribution shift (Ben-David et al. 2006). Unsupervised domain adaptation (UDA) is proposed to adapt a source model to a new unlabeled target domain. Most UDA research considers the adaptation from single or multiple sources to a single target (STDA). However, in reality, the target domain can be diverse and include various styles and textures, and the distribution of each class also varies from each target. These steer us to consider a practical yet challenging setting termed as *blended targets domain adaptation (BTDA)*: 1)Adaptation is conducted from one single source to multiple targets. 2)Neither domain labels nor class labels are available on targets and the model should perform well on each target. 3)Label distributions of different targets can be different (label shift). In the following, we first present the analysis of BTDA and discuss limitations of current methods due to these essential issues. Finally, we discuss that domain labels are not *directly* necessary for BTDA and propose the category-oriented mutual conditional domain adaptation (MCDA), which also generalizes to common settings.

There are two practical issues for distributional alignment in BTDA: 1)Diverse styles and textures of blended targets. 2)Label shift of various targets. These induced our key observation that categorical feature space in BTDA is hybrid and unstructured as shown in Figure 1. Features of different classes in the blended targets are pervasive and do not form a well-clustered structure. To analyze it, we conduct t-SNE for feature space under BTDA in the left. Besides, we also uniformly sample and calculate K nearest neighbors (KNN) of each class center under STDA and BTDA. The result in the middle shows that the number of samples within the same class in STDA is more than that of BTDA. This indicates that the cluster structure of BTDA is not well formed compared to STDA which corresponds to the hybrid categorical feature space in t-SNE visualization. This *weakens* the cluster assumption (Chapelle and Zien 2005) that serves as the necessary condition of many adaptation methods (Tachet des Combes et al. 2020; Shu et al. 2018; Tang, Chen, and Jia 2020; Yang et al. 2021). Further, it motivates our analytical perspective from both categorical distribution shift and biased classifier for BTDA.

Current UDA methods yield sub-optimal performance in BTDA due to these. Concretely, methods based on the *covariate shift* assumption and aligning marginal distributions (Tzeng et al. 2017; Ganin and Lempitsky 2015; Shen et al. 2018) are inevitable to increase the joint error of optimal hypotheses under the label shift (Wu et al. 2019; Tachet des Combes et al. 2020). BTDA worsens the situation with diverse target domains and more serious imbalance and label shift issues. Some theories further propose conditional alignment formulation to avoid the joint error issue. (Tachet des Combes et al. 2020; Jiang et al. 2020) implicitly align conditional distribution by aligning reweighted marginal distributions that still need cluster assumption. Other methods use the target pseudo labels to model and align conditional distributions through class centroids (Pan et al. 2019; Tanwisuth et al. 2021; Singh 2021), task-oriented classifiers (Zhang et al. 2019; Saito et al. 2019), and the conditional discriminator (Long et al. 2018). These effec-

* Corresponding author

tive STDA methods produce limited performance under the hybrid feature space in BTDA. Centroid and general adversarial methods may not model distributions well. The biased classifier and clustering labeling algorithm also generate noisy labels in this situation.

Recent multi-target domain adaptation (MTDA) methods produce impressive results by generally inferring or utilizing the domain level information and then conducting STDA methods. Some methods train separated models for each target which is not efficient in practice (Saporta et al. 2021; Isobe et al. 2021; Nguyen-Meidine et al. 2021; Saporta et al. 2021). Other methods utilize graph neural networks with co-teaching (Roy et al. 2021a), disentanglement methods (Gholami et al. 2020) and meta-clustering (Chen et al. 2019). These require domain labels and lack consideration for the imbalance and the hybrid target feature space in BTDA.

In this paper, we address two intrinsic issues of BTDA: 1)Domain labels. 2)Hybrid categorical feature space. First, our analysis shows that domain labels are not *directly* necessary for BTDA only if the categorical distributions of various domains are sufficiently aligned even facing the imbalance and the label shift. However, categorical alignment requires labels. The hybrid categorical feature space in BTDA raises practical issues in modeling categorical distributions and producing reliable pseudo labels. Considering these, we design techniques to explicitly model and align categorical distributions $P(Z|Y)$ of various domains and simultaneously correct the biased classifier $P(Y|Z)$ among diverse targets to enhance pseudo labels.

Practically, this motivates two designs on $P(Z|Y)$ and $P(Y|Z)$. Firstly, for modeling and aligning $P(Z|Y)$, current methods such as prototype (Pan et al. 2019; Tanwisuth et al. 2021) and kernel methods (Wang et al. 2020) inferiorly model conditional distributions of unstructured data features in the hybrid feature space in Figure 1. Leveraging the distribution modeling ability of GAN (Arora et al. 2017; Goodfellow et al. 2014), we propose an uncertainty-guided categorical domain discriminator. We encode categorical distributions within the same semantic space to *explicitly* model and *directly* align $P(Z|Y)$ of various domains. Since the discriminator is supervised with source and noisy target labels, we adopt uncertainty to guide it to gradually learn and align categorical distributions. Secondly, to correct the biased classifier for reliable pseudo labels during adaptation, we first adopt balanced sampling on the source data and then utilize the low-level features in convolution neural networks (CNN) to augment the source features with diverse target styles to reduce domain dependent information and balance the classifier training on target classes. Our method shows that one single labeled source can still be augmented with multiple targets to rectify the classifier during adaptation by leveraging the prior of low-level features in CNN.

In summary, our contributions are as follows: 1)We demonstrate that the adaptation can be well-achieved without domain labels in BTDA if categorical distributions are sufficiently aligned even facing the imbalance and label shift. 2)We propose the mutual conditional alignment to directly minimize conditional distributions and simultaneously correct the biased classifier. 3)Practically, to address the hybrid feature space of BTDA, we design an uncertainty-guided categorical domain discriminator to explicitly model and align categorical distributions, and utilize low-level features to mitigate the bias of the classifier on blended targets. Our method achieves state-of-the-art in BTDA even compared with methods using domain labels, especially under the label shift, and in STDA with DomainNet.

Related Works

Single Target UDA (STDA): STDA is a typical setting that adapts single or multiple sources into one target. Generally, the research includes four categories. One branch minimizes the explicit statistical distance such as Maximum Mean Discrepancy (MMD) to mitigate the domain distribution shift (Long et al. 2015, 2017; Venkateswara et al. 2017; Tzeng et al. 2014; Shen et al. 2018; Lee et al. 2019; Xu et al. 2019; Montesuma and Mboula 2021). The second branch leverages the adversarial training to implicitly minimize the domain discrepancy through GAN (Ganin and Lempitsky 2015; Tzeng et al. 2017; Zhang et al. 2019) or entropy minimization (Pan et al. 2020; Vu et al. 2019). The third one utilizes the self-training with the target pseudo labels to train the source model (Liu, Wang, and Long 2021; French, Mackiewicz, and Fisher 2017). The fourth one utilizes image translation techniques to mitigate the semantic irrelevant gap (Sankaranarayanan et al. 2018; Roy et al. 2021b; Kim and Byun 2020; Yang et al. 2020a). However, these methods produce limited performance in BTDA. The serious imbalance and label shift issues in blended targets cause a serious incremental error of classifiers (Wu et al. 2019), and the hybrid target feature space also yields noisy pseudo labels and calibration issues (Mei et al. 2020), which deteriorates the self-training and conditional alignment methods.

Multi-Target UDA (MTDA): transfers the knowledge from a single source to multiple targets. MTDA is recently studied in both classification (Gholami et al. 2020; Nguyen-Meidine et al. 2021; Chen et al. 2019; Roy et al. 2021a; Yang et al. 2020b) and semantic segmentation (Saporta et al. 2021; Isobe et al. 2021). One common approach is to disentangle domain information from multiple targets by adversarial learning and adapt each target with a separated network (Saporta et al. 2021; Gholami et al. 2020). AMEAN (Chen et al. 2019) first clusters blended targets into sub-clusters and adapt the source with each cluster. CGCT (Roy et al. 2021a) uses graph convolution network (GCN) for feature aggregation and uses GCN classifier and source classifier for co-teaching. Differently, our method does not require any domain label and conducts BTDA in a united network which is scalable and efficient. Besides, our model considers the hybrid categorical feature space and is robust under the imbalance and label shift in BTDA.

Methodology

We present our analysis of BTDA and discuss the proposed mutual conditional domain adaptation (MCDA) framework including: explicit categorical adversarial alignment, uncertainty-guided discriminative adversarial training, and low-level feature manipulation for the classifier correction.

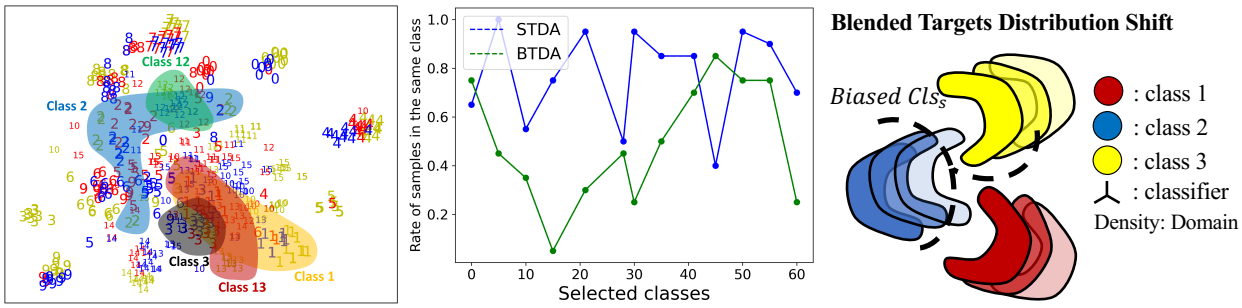

Figure 1: Left: t-SNE for hybrid categorical feature space of BTDA where features of various classes are pervasive and unstructured. The color indicates the domain and the digit indicates the class. Middle: the sample rate of the same class for each class center's K nearest neighbors. All data are collected from Office-Home (ResNet-50). Right: BTDA distribution shift where features are unstructured and the classifier is biased.

Notation. Let us denote the input-output space $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ represents the image space and $\mathcal{Y}$ represents the label space. The labeled source domain is denoted as $\mathcal{S} = \{x_i^s, y_i^s\}_{i=1}^{|S|}$ and each unlabeled target is denoted as $\mathcal{T}_j = \{x_i^{t_j}\}_{i=1}^{|T_j|}$. Both source and target domains are i.i.d sampled from some distribution $P_{\mathcal{S}}(X,Y)$ and $P_{\mathcal{T}_j}(X,Y)$. For the model, we denote the feature extractor $g : \mathcal{X} \to \mathcal{Z}$ and the classifier $h : \mathcal{Z} \to \mathcal{Y}$. The error rate of the model on source $\mathcal{S}$ and target $\mathcal{T}_j$ are ϵ_S an ϵ_{T_j}, and the blended target error rate is evaluated as $\epsilon_T = \frac{1}{K}\sum_j \epsilon_{T_j}$.

MCDA is a unified framework that adapts a single source to the blended targets such that the model performs well on each single target even facing the label distribution shift across various targets. i.e., $P_{\mathcal{S}}(Y) \neq P_{\mathcal{T}_j}(Y)$ and $P_{\mathcal{T}_j}(Y) \neq P_{\mathcal{T}_m}(Y)$. As proved in (Tachet des Combes et al. 2020), minimizing marginal distribution shift of $P_{\mathcal{S}}(X)$ and $P_{\mathcal{T}}(X)$ can arbitrarily increase the target error ϵ_T due to the label shift, which fails the adaptation. The situation becomes worse in BTDA since each target can have a different $P_{\mathcal{T}_m}(Y)$. In that, we are interested to have a bound such that each term of it can be independently minimized as much as possible, and that is better irrelevant with domain labels.

Blended Error Decomposition Theorem. Inspired by the generalized label shift (GLS) theorem in (Tachet des Combes et al. 2020), we intend to align the conditional distributions of each class $P(Z|Y)$ within each target to the same class in the source domain on the feature space $\mathcal{Z}$.

Theorem 1 *For any classifier $\hat{Y} = (h \circ g)(X)$, the blended target error rate is*

$$\|\epsilon_S - \frac{1}{K}\sum_j^K \epsilon_{T_j}\| \leq \frac{1}{K}\sum_j^K \|P_{\mathcal{S}}(Y) - P_{\mathcal{T}_j}(Y)\|_1 BER_{P_{\mathcal{S}}}(\hat{Y}\|Y) + 2(c-1)\Delta_{BTCE}(\hat{Y}). \tag{1}$$

$$BER_{P_{\mathcal{S}}}(\hat{Y}\|Y) = \max_{j\in[k]} P_{\mathcal{S}}(\hat{Y} \neq Y|Y=j) \tag{2}$$

$$\Delta_{BTCE}(\hat{Y}) = \frac{1}{K}\sum_j^K \max_{y\neq y'\in\mathcal{Y}^2} |P_{\mathcal{S}}(\hat{Y}\neq Y|Y=y) - P_{\mathcal{T}_j}(\hat{Y}\neq Y|Y=y)| \tag{3}$$

where $\|P_{\mathcal{S}}(Y) - P_{\mathcal{T}_j}(Y)\|_1$ represents the L_1 distance of label distributions between the source and each target and is a constant only depending on the data, $BER_{P_{\mathcal{S}}}(\hat{Y}\|Y)$ is the classification performance only related with the source domain. $\Delta_{BTCE}(\hat{Y})$ measures the conditional distribution discrepancy of each class between the source and each target. In this sense, we only need to minimize the $\Delta_{BTCE}(\hat{Y})$, which is equivalent to minimize the discrepancy between $P_{\mathcal{S}}(Z|Y=y)$ and $P_{\mathcal{T}_j}(Z|Y=y)$.

Key Differences. First, different from (Tachet des Combes et al. 2020), we argue that the theorem 3.3: clustering structure assumption in (Tachet des Combes et al. 2020) is a strong assumption in BTDA because each target has a different cluster structure $Z_{\mathcal{T}_j}$ in feature space under the pre-trained feature extractor $g_{\mathcal{S}}$, which induces hybrid categorical feature space and different decision boundaries for different target $\mathcal{T}_j$ as illustrated in Figure 1. When blended together, the cluster of class a in $\mathcal{T}_m$ may overlap with the cluster of class b in $\mathcal{T}_n$. Consequently, the sufficient condition for GLS may not hold. Thus, calculating class ratios and aligning reweighted marginal distributions in (Tachet des Combes et al. 2020) do not induce GLS to align the semantic conditional distributions. Second, when the number of classes $|\mathcal{Y}|$ is large, solving a quadratic problem to find class ratios requires $O(|\mathcal{Y}|^3)$ time which is not efficient and accurate. We do not calculate class ratios. Finally, we do not make any assumption to satisfy GLS. Instead, we adopt the general bound in equation 1 and design model to directly minimize the conditional JS divergence $\mathcal{D}_{JS}(P_{\mathcal{S}}(Z|Y=y)\|P_{\mathcal{T}_j}(Z|Y=y))$ to enforce it. However, aligning conditional distribution requires accurate pseudo labels. This motivates us to develop a mutual conditional alignment system to align $P(Z|Y)$ and $P(Y|Z)$ simultaneously. Besides, since we only use the class label, the domain label is unnecessary, which suits the BTDA setting.

Explicit Categorical Adversarial Alignment

Our motivation is to *explicitly* model and *directly* align the categorical JS divergence $\mathcal{D}_{JS}(P_{\mathcal{S}}(Z|Y=y)\|P_{\mathcal{T}_j}(Z|Y=y))$ between the source and each target under the hybrid feature space. Current categorical alignment methods utiliz-

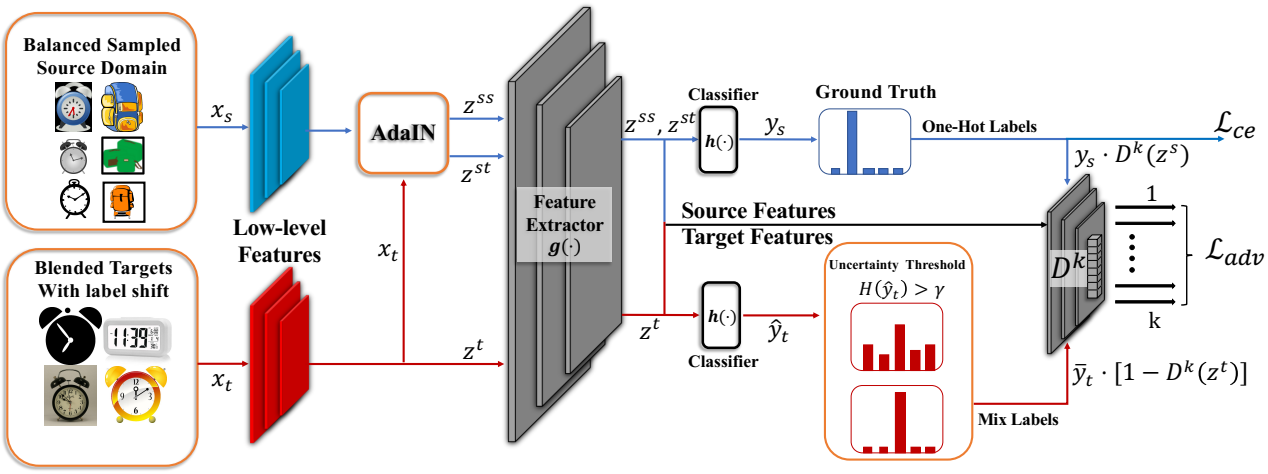

Figure 2: The framework of MCDA. The source data utilizes balanced sampling for training the categorical discriminator and is augmented with blended target styles to train the classifier. The target data is randomly sampled, and the predicted pseudo labels with low uncertainty are converted to one-hot labels to train the categorical domain discriminator.

ing task-classifiers, prototypes, and conditional discriminator (Zhang et al. 2019; Saito et al. 2019; Long et al. 2018) may not represent conditional distributions well in the hybrid categorical feature space in BTDA.

Leveraging the distribution modeling ability of GAN, we intend to encode categorical distributions of various domains into the same semantic space to explicitly model categorical distributions for optimization. Inspired by DANN (Ganin and Lempitsky 2015), we augment the last layer of a general domain discriminator D into the number of classes k, and each logit of such a categorical domain discriminator D^k is followed by a sigmoid function to predict the probability of a feature belonging to the source or target domain conditional on the corresponding class. Each logit behaves as a single GAN to minimize the discrepancy of JS divergence of a specific class $P_{\mathcal{S}}(Z|Y=y)$ and $P_{\mathcal{T}_j}(Z|Y=y)$. To make each logit corresponds to one class in D^k, we feed one feature $g(x_i)$ into D^k and get the prediction $d_i \in R^k$. Then we use the corresponding *one-hot label* $y_i \in \{0,1\}^k$ to only activate the corresponding logit to compute adversarial loss by $y_i \cdot d_i$. To achieve this, we use pseudo target labels and design a strategy to make the categorical adversarial alignment and pseudo label refinement reinforce each other. Then we formulate the optimization as follows

$$\mathcal{L}_{adv}(g, D^k) = \frac{1}{n_s}\sum_{i=1}^{n_s} y_i \cdot log[D^k(g(x_i^s))] + \frac{1}{n_t}\sum_{j=1}^{n_t} \bar{y}_j \cdot log[1 - D^k(g(x_j^t))], \tag{4}$$

where y_i represents the one-hot true labels of the source and $\bar{y}_j$ represent the mix of soft and one-hot pseudo labels of the target. We discuss this in detail in the next section.

Uncertainty Guided Discriminative Adversarial Training

To train a discriminative categorical domain discriminator D^k, we require the one-hot true labels of source and blended targets. However, since the initial target labels are noisy, we design the uncertainty-guided training strategy for our categorical domain discriminator. We start with soft target labels and then gradually convert soft target labels with low uncertainty into one-hot encoding as training goes by. We use the entropy as the metric of the uncertainty of each sample and select the samples based on a threshold γ.

$$H(x_j) = -\sum_{k=1}^{K} \hat{y}_{j,k} \cdot log(\hat{y}_{j,k}) \tag{5}$$

$$\bar{y}_j = \begin{cases} \{0,1\}^k, & if H(x_j) < \gamma \\ \hat{y}_j, & otherwise \end{cases} \tag{6}$$

In the early stage, the entropy of soft pseudo labels on the target domain is large so that each digit of D^k is assigned with a similar probability mass. D^k cannot discriminate different classes and behaves as the general discriminator D in DANN since all logits share the same semantics. As the training goes, the entropy of target pseudo labels will decrease, and the labels will become more discriminative owing to the distribution alignment. At the same time, the discriminative target labels will also train D^k to distinguish different categories and further align the categorical distributions, which forms a mutually reinforced process.

Source-Only Balanced Adversarial Training

We expect our model to be robust under the label shift across various domains such as in a case shown in Figure 3. Equation 1 indicates that the label shift only influences the classification error $BER_{P_{\mathcal{S}}}$ but does not influence the major distribution discrepancy Δ_{BTCE}. It indicates that only if Δ_{BTCE} is small enough, the model is robust to label shift even in the blended domains.

So, we focus on training D^k since the class imbalance will lead to bias to the training of D^k. Hence, D^k cannot distinguish different classes and align distributions biased towards the majority classes, which will ruin the categorical

distribution alignment. To train a balanced D^k, we propose to only conduct balanced sampling for the source domain rather than on both domains as in (Jiang et al. 2020) for two reasons: 1)We only have true labels on the source domain, balanced sampling based on the hard target pseudo labels may introduce errors and bias because initially target pseudo labels are inaccurate. Filtering out confident target pseudo labels may rule out some classes, which exacerbates the class imbalance issue. 2)With conventional double-side balanced sampling, target pseudo labels are only updated every epoch. Instead, mixed target pseudo labels can be updated online with the distribution alignment, which is more beneficial for adaptation. We demonstrate the robustness and efficiency in the Experiments section.

Low-Level Feature for Classifier Correction

We intend to correct the biased classifier $P(Y|Z)$ from a single source to blended targets during the adaptation process. This improves the pseudo label accuracy on blended targets during the adaptation process and further facilitates the training of our categorical domain discriminator. Inspired by the research on low-level features on CNN, we utilize the low-level features of CNN that mainly represent the style and background of images to project blended target styles into the source for correcting the classifier. Denoting the low-level feature maps $z \in \mathbb{R}^{D\times H\times W}$ where D represents the channel and H, W represents the spatial size, leveraging the AdaIN, we have augmented features z^{st} with source content and target style as below:

$$\mu_t = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} z^t \tag{7}$$

$$\sigma_t = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(z^t-\mu_t)^2+\epsilon} \tag{8}$$

$$z^{st} = AdaIN(z^s, z^t) = \sigma_t(\frac{z^s-\mu_s}{\sigma_s}) + \mu_t \tag{9}$$

Compared with previous image translation methods, our method does not need to generate specific images, which is efficient in practice. Besides, considering the diversity and imbalance of blended targets, our method achieves the correction on two sides: 1)Since the source is evenly resampled on class, the augmented feature z^{st} with source content is balanced on semantic classes, which forms a balanced classifier for inference. 2)The augmented z^{st} with diverse target styles mitigate the domain irrelevant information. This regularizes the hybrid categorical feature space in BTDA and make the cluster assumption more practical.

Overall Objective: Eventually, the final loss function consists of the categorical adversarial loss and the classification loss of various domains. Note that h' indicates the networks excluding the shallow layers.

$$\mathcal{L}_{cls}(g,h) = \frac{1}{n_s}\sum_{i=1}^{n_s} l_{ce}(g\circ h(x_i^s), y_i^s) + \frac{1}{n_s}\sum_{i=1}^{n_s} l_{ce}(g\circ h'(z_i^{st}), y_i^s) \tag{10}$$

$$\min_{g,h}\max_{D}\mathcal{L} = \mathcal{L}_{cls}(g,h) + \mathcal{L}_{adv}(g, D^k) \tag{11}$$

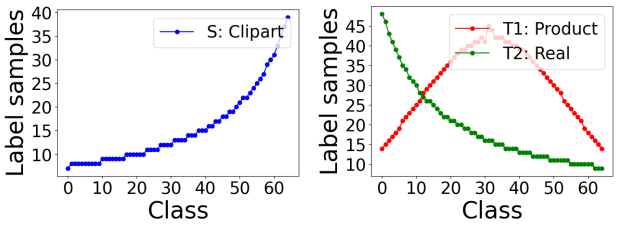

Figure 3: Label distribution shift of Office-Home-LMT.

Experiments

Datasets. We evaluate our method based on standard BTDA tasks (Chen et al. 2019; Roy et al. 2021a): Office-31 (Saenko et al. 2010), Office-Home (Venkateswara et al. 2017), DomainNet (Peng et al. 2019), and a specialized dataset Office-Home-LMT for label shift in BTDA. Similar to Office-Home-RS-UT (Jiang et al. 2020), we use **Cl**, **Pr** and **Rw** to resample two reverse long-tailed distributions and one Gaussian distributions for each of them for BTDA with label shift. For evaluation, we use one domain as the source and the rest as blended targets. The performance is evaluated as the mean accuracy of all target domains. We show a concrete example of **Ar** as the source in the label shift setting in Figure 3.

Baselines and Implementations. We compare our method with previous state-of-the-arts in standard BTDA and that with label shift, i.e., MTDA (Nguyen-Meidine et al. 2021), CGCT (Roy et al. 2021a), MDDIA (Jiang et al. 2020), CST (Liu, Wang, and Long 2021), and SENTRY (Prabhu et al. 2021). For comparison with BTDA under label shift, we also combine the sampling strategies in MDDIA with CGCT. The detailed summary of comparison methods is in the supplementary. We followed the implementations in (Junguang Jiang, Baixu Chen, Bo Fu, Mingsheng Long 2020). For all datasets, we use SGD optimizer with learning rates $\eta_0 = 0.01$, $\alpha = 10$, and $\beta = 0.75$. We set the uncertainty threshold $\gamma = 0.05$ for all datasets. Since CST and SENTRY use AutoAugment for data augmentation, we set the number of transformations $N = 1$ and the transform severity $M = 2.0$ in AutoAugment (Lim et al. 2019) for a fair comparison.

Standard BTDA. We summarize the standard BTDA in Table 1. Our method outperforms comparison methods with a clear margin. Concretely, our method outperforms the latest BTDA methods (e.g., AMEAN and CGCT) by 1.4% on Office-31, 4.6% on Office-Home, and 2.2% on DomainNet. Moreover, even compared with methods utilizing ground truth domain labels, our method can still outperform them by 0.8% on Office-31 and 1.3% on Office-Home. The results validate our argument that categorical distribution alignment overwhelms in BTDA, and echos the theoretical intuition from equation 1 that proper domain alignment is achievable even without domain labels in BTDA.

BTDA with Label Shift. We analyzed the essential label shift influence on BTDA and summarized results of the specialized Office-Home-LMT in Table 2. The label distribution of each domain is different from each other. Since CST and SENTRY essentially require extra augmented data, we

Methods	Office-31				Office-Home					Methods	DomainNet						
	A	D	W	Avg.	Ar	Cl	Pr	Rl	Avg.		Cli	Inf	Pai	Qui	Rea	Ske	Avg.
Source	68.6	70.0	66.5	68.4	47.6	42.6	44.2	51.3	46.4	Source	25.6	16.8	25.8	9.2	20.6	22.3	20.1
DAN	79.5	80.3	81.2	80.4	55.6	56.6	48.5	56.7	54.4	SE	21.3	8.5	14.5	13.8	16.0	19.7	15.6
DANN	80.8	82.5	83.2	82.2	58.4	58.1	52.9	62.1	57.9	MCD	25.1	19.1	27.0	10.4	20.2	22.5	20.7
CDAN	93.6	80.5	81.3	85.1	59.5	61.0	54.7	62.9	59.5	CDAN	31.6	27.1	31.8	12.5	33.2	35.8	28.7
JAN	84.2	74.4	72.0	76.9	58.3	60.5	52.2	57.5	57.1	DADA	26.4	20.0	26.5	12.9	20.7	22.8	21.5
AMEAN	90.1	77.0	73.4	80.2	64.3	65.5	59.5	66.7	64.0	MCC	33.6	30.0	32.4	13.5	28.0	35.3	28.8
CGCT	**93.9**	85.1	85.6	88.2	67.4	68.1	61.6	68.7	66.5	CGCT	36.1	33.3	35.0	10.0	**39.6**	39.7	32.3
Ours	92.4	**87.7**	**88.8**	**89.6**	**71.7**	**72.8**	**68.0**	**71.7**	**71.1**	Ours	**37.5**	**37.3**	**36.6**	**17.8**	36.1	**41.4**	**34.5**
MTDA†	87.9	83.7	84.0	85.2	64.6	66.4	59.2	67.1	64.3	-	-	-	-	-	-	-	-
DCL†	92.6	82.5	84.7	86.6	63.0	63.0	60.0	67.0	64.1	DCL†	35.1	31.4	37.0	20.5	35.4	41.0	33.4
DCGCT†	93.4	86.0	87.1	88.8	70.5	71.6	66.0	71.2	69.8	DCGCT†	37.0	32.2	37.3	19.3	39.8	40.8	34.4

Table 1: Accurary (%) of BTDA on Office-31, Office-Home (ResNet-50), and DomainNet (ResNet-101). Best results in Bold. Each domain represents the source and the rest domains are blended as the target. The accuracy is the mean of accuracies of all domains in the blended target. †indicates methods using domain labels.

Methods	Clipart	Product	Real	Avg.
Source	42.3	47.6	50.3	46.7
BSP	51.5	52.9	57.4	54.0
CDAN	50.5	53.2	56.3	53.3
DAN	51.0	49.2	56.8	52.3
JAN	51.4	50.1	57.0	53.2
DANN	46.6	50.4	53.3	50.1
ADDA	45.0	49.7	52.8	49.2
MCD	40.2	48.6	52.3	47.0
MDD	43.7	56.0	57.8	52.5
MDDIA	61.9	58.2	63.2	61.1
CGCT	53.7	51.5	52.0	52.4
CGCT+bal	57.1	53.0	56.8	55.7
Ours	**68.0**	**62.3**	**67.5**	**65.9**
CST(aug)	58.3	57.4	63.4	59.7
SENTRY(aug)	65.6	63.5	65.9	65.0
Ours(aug)	**69.1**	**66.2**	**68.9**	**68.1**
Ours(oracle)	98.9	98.3	98.2	98.5
S+T	99.7	99.8	99.8	99.8

Table 2: Accurary (%) of Blended-Office-Home-LMT (ResNet-50). *aug*: using 1 extra augmented data with RandAug (Cubuk et al. 2020). *bal*: using balanced sampling. *oracle*: discriminator trained with true source and target labels. *S+T*: supervised learning on source and target.

add 1 augmented data for each sample and evaluate CST, SENTRY, and ours in the same setting (*aug*). Our method outperforms the label shift UDA method MDDIA by 4.8% and SENTRY by 3.1%. Compared with latest BTDA method CGCT, we get an improvement of more than 12%. We also equip CGCT with balanced sampling strategy in MDDIA (i.e., CGCT+bal) whose result is still inferior to ours.

The result first demonstrates the label shift in BTDA seriously impedes the adaptation, especially for the marginal alignment methods. Second, it validates our proposition in theorem 1 that if categorical distribution $\Delta_{BTCE}(Y)$ can be properly minimized, the label shift only reweights the classification error $BER_{P_S}(\hat{Y}\|Y)$, which is relatively small. Besides, our method does not require balanced sampling on target pseudo labels for every epoch, which can be trained and updated online. The extra data augmentation (e.g., RandAug) is not essentially necessary in the algorithm design.

mix	bal	flip	Office-Home					DomainNet			
			Art	Clip	Prod	Real	Avg.	Real	Info	Pain	Avg.
✓			66.0	67.3	64.5	70.9	67.2	32.0	31.0	33.4	32.1
	✓		70.6	72.5	67.4	69.9	70.1	34.3	33.5	34.9	34.2
		✓	65.0	66.5	65.6	71.0	67.0	34.6	30.7	32.8	32.7
✓	✓		70.2	**73.2**	67.0	70.6	70.3	35.7	35.7	36.1	35.8
✓	✓	✓	**71.7**	72.8	**68.0**	**71.7**	**71.1**	**36.1**	**37.3**	**36.6**	**36.7**

Table 3: Ablations on Office-Home and the selected three domains on DomainNet. *mix*: mix labeling strategy; *bal*: balanced sampling strategy; *flip*: low-level features.

STDA. We also validate the generalization ability of our method in STDA (i.e., Office-Home and DomainNet). We compare our method with SRDC (Tang, Chen, and Jia 2020) which considers cluster structures in STDA, and MDDIA (Jiang et al. 2020) which uses balanced sampling on both source and target domains. Our method achieves 72.4% on Office-Home, and 35.2% on DomainNet which outperforms the previous state-of-the-art method MDD+SCDA (Li et al. 2021) by 1.9%. Please refer to the supplementary for details.

Ablation and Analysis

We present ablations of MCDA in Table 3 on Office-Home and three domains of DomainNet. Each proposed module contributes to the improvement of the final performance.

Effectiveness of Uncertainty for Guiding Adversarial Training. To validate the uncertainty and mixed labels to train a categorical discriminator D^k that mutually reinforces

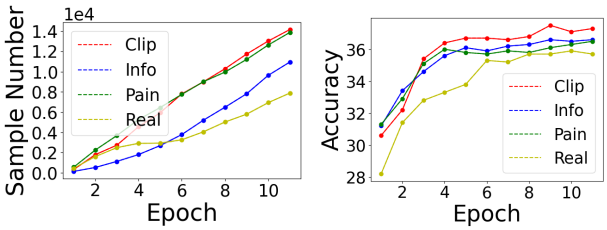

Figure 4: Samples below uncertainty threshold and pseudo label accuracy during training process on DomainNet.

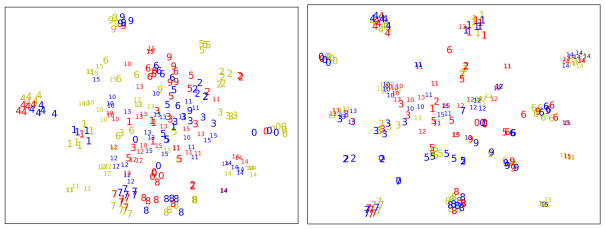

Figure 5: t-SNE feature visualization on Office-Home-LMT with *Clipart* as the source. 15 classes are sampled for conciseness. Color represents the domain and the digit represents the class. *left*: sourceonly, *right*: MCDA.

with target pseudo labels. We show the number of filtered samples below the uncertainty threshold and the corresponding pseudo label accuracy in Figure 4. The results of DomainNet show that during the training, the uncertainty of samples gradually decreases, and more samples pass the threshold. Meanwhile, the accuracy of pseudo labels increases, which justifies our motivation.

Robustness of Uncertainty Threshold. We validate the robustness of our model under various uncertainty thresholds λ in Table 4 for both standard BTDA and BTDA with label shift in Table 4. For main experiments in Table 1 and 2, we set λ as 0.05. The performance is stable when λ is selected from 0.03 to 0.07. The results of Art on Office-Home have a fluctuation range of 0.4%, and results of Clipart on Office-Home-LMT have a fluctuation range of 1.5%. These demonstrate the stability and generality of our model for the standard BTDA and label shift setting.

Verification of Error Theorem under Label Shift. We verify our error theorem in equation 1 that if conditional distributions $\Delta_{BTCE}(\hat{Y})$ is sufficiently minimized, the model is robust under label shift in BTDA since the reweighted source error $\|P_{\mathcal{S}}(Y) - P_{\mathcal{T}_j}(Y)\|_1 BER_{P_{\mathcal{S}}}(\hat{Y}\|Y)$ is relatively small. In Table 2, the *oracle* are results when the discriminator is trained with true labels of source and target but the classifier is trained only with source labels. In this case, the categorical discriminator is trained to minimize categorical distributions as much as possible under ground truth supervision. The *S+T* are results where the classifier is trained with both source and target labels. These two results approximate to each other which verifies the theorem.

Essential of Domain Labels. Our theoretical formulation in equation 1 does not require domain labels to minimize

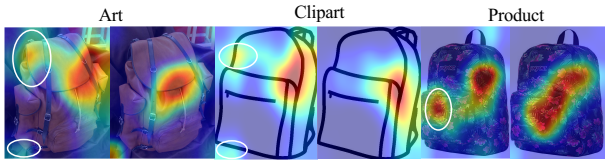

Figure 6: CAM feature response maps on Office-Home-LMT. The left pervasive one is from sourceonly model while the right class-discriminative one is from MCDA.

Thresholds	0.01	0.03	0.05	0.07	0.09
Art	71.9	71.9	71.7	72.3	71.8
Clipart	66.5	67.4	68.0	66.9	66.4

Table 4: Accuracy (%) of different uncertainty thresholds for *Art* in Office-Home and *Clipart* in Office-Home-LMT for BTDA (ResNet-50).

the target error rate in BTDA. The bound is mainly related to the categorical distribution constraint. In comparison with methods (i.e., CGCT (Roy et al. 2021a) and DCL (Nguyen-Meidine et al. 2021)) using domain labels † in Table 1, our method outperforms previous state-of-the-arts by 0.8% on Office-31, 1.3% on Office-Home, and 0.1% on DomainNet even without domain labels. This validates our proposition that adaptation can be done without domain labels in BTDA if categorical distributions is sufficiently aligned.

Feature Visualization. To show our method learns a regular and meaningful categorical feature space, we visualize features of last convolution layer with t-SNE in Figure 5 and CAM (Selvaraju et al. 2017) in Figure 6. The t-SNE visualization further shows that the sourceonly model generates a hybrid feature space while MCDA produces a more class-discriminative feature space, which corroborates our observation on the cluster assumption and categorical alignment on BTDA. The CAM results shows feature response maps of the sourceonly model is pervasive while those of MCDA are more category-discriminative. This validates that MCDA make the classifier learns more task-relevant features and achieve better categorical alignment. More visualization results are discussed in the supplementary.

Conclusion

In this paper, we demonstrate that adaptation can be well achieved without domain labels for BTDA only if the categorical distributions are sufficiently aligned, even facing the cross-domain label shift. Following this, we present the mutual conditional domain adaptation framework. We explore the uncertainty-guided mechanism and source-only balanced sampling strategy to train a categorical domain discriminator for efficiently modeling categorical distributions in BTDA. And we explore low-level features to correct the biased classifier. Extensive experimental results demonstrate the state-of-the-art performance of the framework in single target DA and BTDA tasks under label shift.

Acknowledgments

We appreciate constructive feedback from anonymous reviewers and meta-reviewers. This work is supported by Natural Sciences and Engineering Research Council of Canada (NSERC), Discovery Grants program.

References

Arora, S.; Ge, R.; Liang, Y.; Ma, T.; and Zhang, Y. 2017. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, 224–232. PMLR.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19.

Chapelle, O.; and Zien, A. 2005. Semi-supervised classification by low density separation. In *International workshop on artificial intelligence and statistics*, 57–64. PMLR.

Chen, Z.; Zhuang, J.; Liang, X.; and Lin, L. 2019. Blending-target domain adaptation by adversarial meta-adaptation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2248–2257.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 702–703.

French, G.; Mackiewicz, M.; and Fisher, M. 2017. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Gholami, B.; Sahu, P.; Rudovic, O.; Bousmalis, K.; and Pavlovic, V. 2020. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*, 29: 3993–4002.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Isobe, T.; Jia, X.; Chen, S.; He, J.; Shi, Y.; Liu, J.; Lu, H.; and Wang, S. 2021. Multi-target domain adaptation with collaborative consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8187–8196.

Jiang, X.; Lao, Q.; Matwin, S.; and Havaei, M. 2020. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In *International Conference on Machine Learning*, 4816–4827. PMLR.

Junguang Jiang, Baixu Chen, Bo Fu, Mingsheng Long. 2020. Transfer-Learning-library. https://github.com/thuml/Transfer-Learning-Library. Accessed: 2022-08-01.

Kim, M.; and Byun, H. 2020. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12975–12984.

Lee, C.-Y.; Batra, T.; Baig, M. H.; and Ulbricht, D. 2019. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10285–10295.

Li, S.; Xie, M.; Lv, F.; Liu, C. H.; Liang, J.; Qin, C.; and Li, W. 2021. Semantic concentration for domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9102–9111.

Lim, S.; Kim, I.; Kim, T.; Kim, C.; and Kim, S. 2019. Fast autoaugment. *Advances in Neural Information Processing Systems*, 32.

Liu, H.; Wang, J.; and Long, M. 2021. Cycle Self-Training for Domain Adaptation. *arXiv preprint arXiv:2103.03571*.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, 97–105. PMLR.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, 2208–2217. PMLR.

Mei, K.; Zhu, C.; Zou, J.; and Zhang, S. 2020. Instance adaptive self-training for unsupervised domain adaptation. In *European conference on computer vision*, 415–430. Springer.

Montesuma, E. F.; and Mboula, F. M. N. 2021. Wasserstein Barycenter for Multi-Source Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16785–16793.

Nguyen-Meidine, L. T.; Belal, A.; Kiran, M.; Dolz, J.; Blais-Morin, L.-A.; and Granger, E. 2021. Unsupervised multi-target domain adaptation through knowledge distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1339–1347.

Pan, F.; Shin, I.; Rameau, F.; Lee, S.; and Kweon, I. S. 2020. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3764–3773.

Pan, Y.; Yao, T.; Li, Y.; Wang, Y.; Ngo, C.-W.; and Mei, T. 2019. Transferrable prototypical networks for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2239–2247.

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

Prabhu, V.; Khare, S.; Kartik, D.; and Hoffman, J. 2021. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8558–8567.

Roy, S.; Krivosheev, E.; Zhong, Z.; Sebe, N.; and Ricci, E. 2021a. Curriculum Graph Co-Teaching for Multi-Target Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5351–5360.

Roy, S.; Siarohin, A.; Sangineto, E.; Sebe, N.; and Ricci, E. 2021b. Trigan: Image-to-image translation for multi-source domain adaptation. *Machine vision and applications*, 32(1): 1–12.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. In *European conference on computer vision*, 213–226. Springer.

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

Sankaranarayanan, S.; Balaji, Y.; Castillo, C. D.; and Chellappa, R. 2018. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8503–8512.

Saporta, A.; Vu, T.-H.; Cord, M.; and Pérez, P. 2021. Multi-Target Adversarial Frameworks for Domain Adaptation in Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9072–9081.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shen, J.; Qu, Y.; Zhang, W.; and Yu, Y. 2018. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-second AAAI conference on artificial intelligence*.

Shu, R.; Bui, H. H.; Narui, H.; and Ermon, S. 2018. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*.

Singh, A. 2021. CLDA: Contrastive Learning for Semi-Supervised Domain Adaptation. *Advances in Neural Information Processing Systems*, 34.

Tachet des Combes, R.; Zhao, H.; Wang, Y.-X.; and Gordon, G. J. 2020. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33: 19276–19289.

Tang, H.; Chen, K.; and Jia, K. 2020. Unsupervised domain adaptation via structurally regularized deep clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8725–8735.

Tanwisuth, K.; Fan, X.; Zheng, H.; Zhang, S.; Zhang, H.; Chen, B.; and Zhou, M. 2021. A Prototype-Oriented Framework for Unsupervised Domain Adaptation. *Advances in Neural Information Processing Systems*, 34.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7167–7176.

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

Venkateswara, H.; Eusebio, J.; Chakraborty, S.; and Panchanathan, S. 2017. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5018–5027.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2517–2526.

Wang, W.; Li, H.; Ding, Z.; and Wang, Z. 2020. Rethink maximum mean discrepancy for domain adaptation. *arXiv preprint arXiv:2007.00689*.

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

Xu, P.; Gurram, P.; Whipps, G.; and Chellappa, R. 2019. Wasserstein distance based domain adaptation for object detection. *arXiv preprint arXiv:1909.08675*.

Yang, J.; An, W.; Wang, S.; Zhu, X.; Yan, C.; and Huang, J. 2020a. Label-driven reconstruction for domain adaptation in semantic segmentation. In *European conference on computer vision*, 480–498. Springer.

Yang, S.; van de Weijer, J.; Herranz, L.; Jui, S.; et al. 2021. Exploiting the Intrinsic Neighborhood Structure for Source-free Domain Adaptation. *Advances in Neural Information Processing Systems*, 34.

Yang, X.; Deng, C.; Liu, T.; and Tao, D. 2020b. Heterogeneous graph attention network for unsupervised multiple-target domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Y.; Liu, T.; Long, M.; and Jordan, M. 2019. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, 7404–7413. PMLR.