

Cross-Domain Gradient Discrepancy Minimization for Unsupervised Domain Adaptation

Zhekai Du¹, Jingjing Li^{*1}, Hongzu Su¹, Lei Zhu², Ke Lu¹

¹University of Electronic Science and Technology of China; ²Shandong Normal University

{zhekaid, hongzus}@std.uestc.edu.cn, lijing117@yeah.net, leizhu0608@gmail.com, kel@uestc.edu.cn

Abstract

Unsupervised Domain Adaptation (UDA) aims to generalize the knowledge learned from a well-labeled source domain to an unlabeled target domain. Recently, adversarial domain adaptation with two distinct classifiers (bi-classifier) has been introduced into UDA which is effective to align distributions between different domains. Previous bi-classifier adversarial learning methods only focus on the similarity between the outputs of two distinct classifiers. However, the similarity of the outputs cannot guarantee the accuracy of target samples, i.e., target samples may match to wrong categories even if the discrepancy between two classifiers is small. To challenge this issue, in this paper, we propose a cross-domain gradient discrepancy minimization (CGDM) method which explicitly minimizes the discrepancy of gradients generated by source samples and target samples. Specifically, the gradient gives a cue for the semantic information of target samples so it can be used as a good supervision to improve the accuracy of target samples. In order to compute the gradient signal of target samples, we further obtain target pseudo labels through a clustering-based self-supervised learning. Extensive experiments on three widely used UDA datasets show that our method surpasses many previous state-of-the-arts.

1. Introduction

Conventional deep learning methods suffer from the challenge of heavy dependency on large-scale labeled data, which is extremely expensive in many real-world scenarios such as medical image analysis. To avoid expensive data annotation, unsupervised domain adaptation (UDA) [30, 19, 18] attempts to transfer a model trained on labeled data collected in the source domain to a similar target domain with unlabeled data. To mitigate the domain shift, one popular paradigm in UDA is to reduce the distribution divergence between domains by minimizing a specific metric [24, 27, 42]. Another widely used paradigm aims to learn

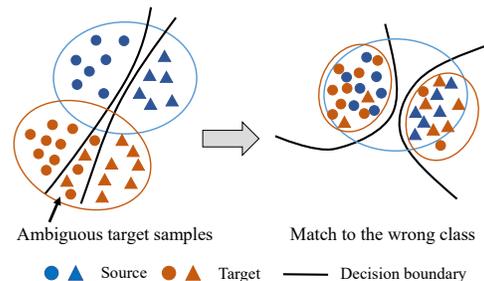


Figure 1. Illustration of the issue in previous bi-classifier adversarial learning. Previous methods only consider the discrepancy between classifiers and neglect the accuracy of the target samples.

domain-invariant feature representations by leveraging the idea of adversarial learning [9], which has achieved remarkable success in the field of UDA recently.

Existing adversarial domain adaptation methods can be implemented in two ways. One way is to apply an extra domain discriminator to distinguish whether a sample comes from the source or the target domain. At the same time, a feature extractor is used to fool the domain discriminator by learning undistinguishable features from input samples [10, 9, 25]. However, these domain adversarial methods neglect the category information of target samples, which may result in deterioration of the feature discriminability [5]. Another adversarial paradigm proposes a within-network adversarial strategy with two classifiers [36, 14]. Through the minimax game between the classifiers and the generator on the cross-classifier outputs discrepancy, the target samples outside the support of the source domain can be detected by the decision boundaries effectively, thus the feature alignment could be established while the discriminability is also preserved.

Although bi-classifier adversarial learning has shown promising performance, methods with this paradigm only focus on the similarity between two distinct classifiers through a discrepancy metric such as ℓ_1 distance [36] and slide wasserstein distance [14]. Here we argue that only considering the discrepancy between classifiers cannot guarantee the accuracy and diversity of classification on target samples, because it is possible that both classi-

*Jingjing Li is the corresponding author.

fiers get wrong results. As shown in Fig. 1, ambiguous target samples may be detected by wrong decision boundaries, which inevitably results in an inaccurate class-wise distribution alignment in subsequent adversarial procedure, although the discrepancy metric, e.g., ℓ_1 and Wasserstein distance between the outputs of two classifiers is small. The main reason is that previous bi-classifier adversarial methods lack the consideration of the accuracy on target samples.

In this paper, we aim to alleviate this issue. One straightforward idea is to label target samples by pseudo labels and fine-tune the model with data of both domains, which has been proven to be effective in UDA [6]. However, directly using hard pseudo labels for supervised learning would lead to error diffusion and converge to the accuracy of pseudo labels. Here we tackle this issue from another perspective, we notice that the gradient discrepancy between source and target samples is related to the accuracy: assuming there is an accurate classifier, the source data and the target data would produce similar gradient signals for updating the classifier. Our key idea is that we wish the loss of two domains to be close to not only the final model but also to follow a similar path to it throughout the optimization. Therefore, we use the gradient signal as a surrogate supervision and make it a proxy to the classification accuracy, yielding a cross-domain gradient discrepancy minimization (CGDM) method for UDA. CGDM employs the gradient discrepancy between source and target samples as an extra supervision. Furthermore, considering that pseudo labels obtained by the source-only classifier may be not accurate enough, we leverage a clustering-based self-supervised method to obtain more reliable pseudo labels for target samples. By aligning gradient vectors, distributions of two domains can be better aligned at the category-level. The main contributions of this work are summarized as follows:

- In order to solve the inaccurate alignment issue in previous bi-classifier adversarial learning, we propose a novel method which explicitly minimizes the discrepancy of gradient vectors produced by the source and the target samples. Notably, we formulate the proposed gradient discrepancy minimization as a generalized learning loss which can be easily applied to other UDA paradigms.
- For computing the gradient of target samples, we employ a clustering-based strategy to obtain more reliable pseudo labels. Then a self-supervised learning based on pseudo labels is conducted to fine-tune the model with both the source data and the target data in order to reduce the number of ambiguous target samples.
- We reformulate the vanilla bi-classifier adversarial framework with above two proposals and conduct extensive experiments on three open large-scale datasets. The experimental results demonstrate the advantage of our method.

2. Related Work

The research line of existing unsupervised domain adaptation can be roughly divided into three branches: sample weighted adaptation, metric learning adaptation and adversarial learning adaptation. **Sample weighted adaptation** aims to reduce the discrepancy between domains by inferring resampling weights of samples in a non-parameter way [38, 41, 17, 20]. **Metric learning** methods try to mitigate the distribution gap directly by minimizing a discrepancy metric. For instance, maximum mean discrepancy (MMD) is a widely used criteria to measure the divergence between different domains in previous work. Deep adaptation networks (DAN) [24] simultaneously minimize the multi-kernel MMD between two domains and the accuracy error on source samples, by which the distributions can be aligned. Joint adaptation networks (JAN) [27] extends DAN by aligning the joint distributions of multiple domain-specific layers using the joint maximum mean discrepancy (JMMD). Besides, some other variants of MMD are also used for more appropriate divergence criteria [26, 21, 13]. In addition, central moment discrepancy (CMD) [42] and maximum density divergence (MDD) [16] are another two criterias to align feature distributions in hidden layers.

Adversarial learning methods learn domain-invariant feature representations following an adversarial paradigm. The adversarial learning of this paradigm can be realized through two strategies. The first way is to employ an additional domain discriminator to distinguish domain-specific features and a feature learner is then employed to learn undistinguishable features to fool the discriminator [9, 37]. Later, some studies [25, 32, 44, 15] suggest to align conditional distributions in feature space to achieve accurate alignment at the category-level. Different from them, our method falls into the second adversarial paradigm, which uses two distinct task-specific classifiers to oppose the generator *w.r.t.* the prediction discrepancy of two classifiers. This paradigm is first used in MCD [36]. SWD [14] further improves the discrepancy metric in MCD by applying the slide Wasserstein discrepancy rather than simple ℓ_1 distance. Recently, CLAN [28] use cosine similarity of the classifier parameters to measure the discrepancy for semantic segmentation. However, accurate alignment in this paradigm cannot be accessed by previous methods.

Pseudo label based methods. Some recent UDA methods use pseudo labeling technique to exploit semantic information of target samples. Zhang *et al.* in [43] directly use pseudo labels as a regularization. Xie *et al.* [39] utilize pseudo labels to estimate class centroids for the target domain and match them to the ones in the source domain. Long *et al.* [25] use pseudo labels predicted by the model to achieve conditional distribution alignment. Zou *et al.* [45] further propose a self-training framework that alternately refines pseudo labels and performs model training. Re-

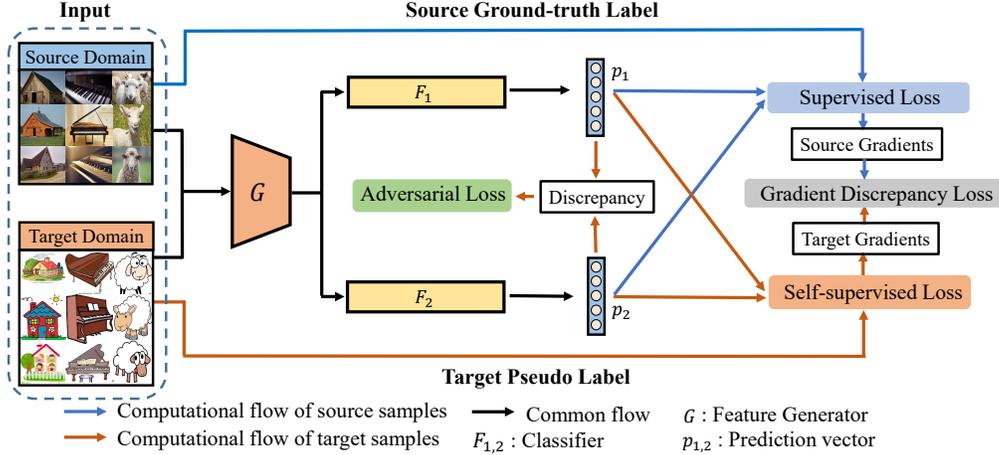


Figure 2. An illustration of our framework. Both source samples and target samples are passed through the generator G and two classifiers F_1 and F_2 . Then the supervised loss is used to minimize the classification error on source samples. We use a self-supervised mechanism to reduce the number of ambiguous target samples. The minimax process on the adversarial loss is used to detect target samples outside the support of the source domain. The gradient discrepancy loss is further used to align distributions and improve the accuracy of target samples.

cently, clustering-based pseudo labeling methods [2, 22, 13] have shown their superiority and successfully applied to domain adaptation. CAN [13] solves the target pseudo labels by performing k-means clustering in feature space for contrastive learning. SHOT [22] leverages a weighted k-means clustering strategy to obtain pseudo labels for self-supervision in the source-free UDA. Our method follows this strategy, instead of directly using pseudo labels for self-training, we propose to minimize the cross-domain gradient discrepancy based on pseudo labels to reinforce the vanilla bi-classifier adversarial learning, which is illustrated in Fig. 2 and elaborated on in subsequent sections.

3. Method

In this section, we first revisit the UDA and the bi-classifier adversarial setting. Then we introduce the cross-domain gradient discrepancy minimization and the self-supervised mechanism respectively. Finally, we report the whole training schema of our CGDM.

3.1. UDA with Bi-Classifier Adversarial Learning

Suppose that we have the source data which consists of n_s labeled samples $\{\mathbf{x}_i^s\}_{i=1}^{n_s}$ and corresponding labels $\{y_i^s\}_{i=1}^{n_s}$ drawn from the source distribution $P(\mathcal{X}^s, \mathcal{Y}^s)$, as well as the target data which consists of n_t unlabeled samples $\{\mathbf{x}_i^t\}_{i=1}^{n_t}$ drawn from the target distribution $P(\mathcal{X}^t)$. UDA aims to obtain a function \mathcal{F} that can predict the category of samples accurately with only source labels accessible and can be applied well to the target domain.

The bi-classifier adversarial learning methods employ a feature generator G which extracts discriminative deep features of raw inputs, and two distinct task-specific classifiers F_1 and F_2 , which are fed with the output of the generator

then produce the prediction probability $p_1(\mathbf{y} | \mathbf{x})$, $p_2(\mathbf{y} | \mathbf{x})$ respectively. Following MCD [36] and SWD [14], we first revisit the standard three-step principle of bi-classifier adversarial learning, which is used as our starting point:

Step 1. Learn G , F_1 and F_2 jointly by minimizing the classification loss $\mathcal{L}_{cls}(\cdot, \cdot)$ on the labeled source samples to reduce the empirical risk over the source distribution, which can be formulated as follows:

$$\min_{\theta_g, \theta_{f_1}, \theta_{f_2}} \mathcal{L}_{cls}(\mathcal{X}^s, \mathcal{Y}^s) = \frac{1}{2n_s} \sum_{i=1}^{n_s} \sum_{n=1}^2 \mathcal{L}_{ce}(F_n(G(\mathbf{x}_i^s)), y_i^s) \quad (1)$$

where $\mathcal{L}_{ce}(\cdot, \cdot)$ denotes the standard cross entropy loss function, θ_g , θ_{f_1} and θ_{f_2} represent the parameters of G , F_1 and F_2 respectively.

Step 2. Frozen the parameters of the generator G and update classifiers F_1 and F_2 to maximize the divergence between their probabilistic outputs on target samples while preserving the classification accuracy on source samples:

$$\min_{\theta_{f_1}, \theta_{f_2}} \mathcal{L}_{cls}(\mathcal{X}^s, \mathcal{Y}^s) - \mathcal{L}_{dis}(\mathcal{X}^t) \quad (2)$$

where $\mathcal{L}_{dis}(\cdot)$ denotes the function that measures the divergence between the probabilistic outputs of two classifiers and can be customized by specific algorithms like ℓ_1 distance [35] and slide wasserstein distance [14].

Step 3. Frozen the parameters of two classifiers F_1 and F_2 , then update the generator G to minimize the divergence between the probabilistic outputs of two classifiers:

$$\min_{\theta_g} \mathcal{L}_{dis}(\mathcal{X}^t) \quad (3)$$

After repeating above steps several times, the model can effectively detect target samples outside the support of the

source domain by decision boundaries and align distributions of two domains to some extent.

3.2. Minimizing Cross-Domain Gradient Discrepancy

As mentioned above, it is difficult to guarantee that the model can classify the target sample with high accuracy if only considering the discrepancy of two classifiers without any other constraints. For instance, we have two predictions $[0.95, 0.03, 0.02]$ and $[0.96, 0.02, 0.02]$ of a target sample, one for each classifier. Although the discrepancy metric of two classifiers is quite small, we can not guarantee it is a good prediction because the vanilla bi-classifier adversarial learning is accuracy-agnostic, the groundtruth label of this sample may be $[0,1,0]$ or $[0,0,1]$. Thus there is a strong incentive to tackle this issue.

Based on the above motivation, we look at the problem from another perspective: if we want to learn a classifier through which all samples from both domains can be classified correctly, then the gradient vector produced by the samples from the source and the target should be similar for learning such a classifier. Therefore, we introduce the gradient similarity metric between two domains. To this end, we first denote the expected gradient over the source and target examples by g_s and g_t respectively. The appropriate gradient of source samples is formulated as follows,

$$g_s = \frac{1}{2} \sum_{n=1}^2 \mathbb{E}_{(\mathbf{x}_i^s, y_i^s) \sim (\mathcal{X}^s, \mathcal{Y}^s)} [\nabla_{\theta_{f_n}} \mathcal{L}_{ce} (F_n (G (\mathbf{x}_i^s)), y_i^s)]. \quad (4)$$

For computing the gradient produced by target samples, the label information is required, however, this is exactly what we want to obtain in the UDA problem. To tackle this awkward situation, we assign pseudo labels to target samples, denoted by \mathcal{Y}^* . Since the pseudo labels may still incorrect, to alleviate the noise of ambiguous target samples which may have incorrect pseudo labels, we use the weighted classification loss based on the prediction entropy of each target sample to compute the gradient. The gradient vector of the target samples is formulated as follows,

$$g_t = \frac{1}{2} \sum_{n=1}^2 \mathbb{E}_{(\mathbf{x}_i^t, y_i^*) \sim (\mathcal{X}^t, \mathcal{Y}^*)} [\nabla_{\theta_{f_n}} \mathcal{L}_{ce}^W (F_n (G (\mathbf{x}_i^t)), y_i^*)]. \quad (5)$$

Here $\mathcal{L}_{ce}^W(\cdot, \cdot)$ is the weighted cross entropy loss function which is formulated as follows,

$$\mathcal{L}_{ce}^W (F_n (G (\mathbf{x}_i^t)), y_i^*) = w_j(\mathbf{x}_i^t) \mathcal{L}_{ce} (F_n (G (\mathbf{x}_i^t)), y_i^*), \quad (6)$$

$$w_j(\mathbf{x}_i^t) = 1 + e^{-E(\delta(F_n(G(\mathbf{x}_i^t)), y_i^*))}, \quad (7)$$

where δ represents the softmax output and $E(\cdot)$ denotes the standard information entropy. At present, we have obtained

the gradient vector of source samples and target samples respectively. We minimize the discrepancy between these two gradient vectors when updating the generator in **step 3** through a gradient discrepancy loss \mathcal{L}_{GD} , here we use cosine similarity to express the discrepancy,

$$\mathcal{L}_{GD} = 1 - \frac{g_s^T g_t}{\|g_s\|_2 \|g_t\|_2}, \quad (8)$$

by which the distributions of the source domain and the target domain are aligned while the semantic information of samples is also considered.

3.3. Self-supervised Learning for Target Samples

In our method, pseudo labels of target samples are indispensable to compute the gradient signal. Pseudo labeling has gained popularity in domain adaptation in recent years. Some previous studies [43, 6, 39] directly incorporate naive pseudo labeling strategy into their methods. However, target pseudo labels that produced by the source model are still unreliable owing to the domain shift. Inspired by DeepCluster [2] and SHOT [22], in this paper, we use a weighted clustering strategy to obtain more reliable pseudo labels. Here we use softmax outputs of two classifiers to weight the samples for obtaining the centroid c_k of the k -th class,

$$c_k = \frac{\sum_{n=1}^2 \sum_{\mathbf{x}_i^t \in \mathcal{X}^t} \delta_k (F_n (G (\mathbf{x}_i^t))) G (\mathbf{x}_i^t)}{\sum_{n=1}^2 \sum_{\mathbf{x}_i^t \in \mathcal{X}^t} \delta_k (F_n (G (\mathbf{x}_i^t)))}, \quad (9)$$

where δ_k means the corresponding k -th element of the softmax output δ . Then pseudo labels could be obtained by the nearest centroid strategy, i.e.,

$$y_i^* = \arg \min_k d (G (\mathbf{x}_i^t), c_k), \quad (10)$$

here d could be any specific distance metric function and we use cosine distance in this paper.

In order to exploit unlabeled samples, we reinforce the **step 1** with the self-supervised learning mechanism to induce the model to learn a discriminative original target distribution and encourage each sample to lie around the correct decision boundary. The weighted classification loss $\mathcal{L}_{cls}^W(\cdot, \cdot)$ for self-supervised learning can be formulated as follows,

$$\mathcal{L}_{cls}^W (\mathcal{X}^t, \mathcal{Y}^*) = \frac{1}{2n_t} \sum_{i=1}^{n_t} \sum_{n=1}^2 \mathcal{L}_{ce}^W (F_n (G (\mathbf{x}_i^t)), y_i^*), \quad (11)$$

Through the above self-supervised process, we make an improvement on the discriminability of the target distribution in **step 1** by fine-tuning the model with both source and target data, so that the samples between two domains could be aligned at category-level subsequently.

3.4. Overall Objective and Optimization Procedure

In this work, we aim to tackle potential drawbacks of the bi-classifier adversarial learning by using self-supervised learning and minimizing the discrepancy of gradient signal produced by the source and the target domain. To sum up, by utilizing the self-supervised loss, the optimization objective of **step 1** could be reformulated as follows,

$$\min_{\theta_g, \theta_{f_1}, \theta_{f_2}} \mathcal{L}_{cls}(\mathcal{X}^s, \mathcal{Y}^s) + \alpha \mathcal{L}_{cls}^W(\mathcal{X}^t, \mathcal{Y}^*), \quad (12)$$

where $\alpha > 0$ is the trade-off parameter and could be adjusted according to the validation set. Following MCD, we use the ℓ_1 distance to estimate the discrepancy between two distinct classifiers in **Step 2**. In **step 3**, we add a constraint on the gradient discrepancy between two domains, thus the final optimization objective becomes

$$\min_{\theta_g} \mathcal{L}_{dis}(\mathcal{X}^t) + \beta \mathcal{L}_{GD} \quad (13)$$

where $\beta > 0$ denotes the trade-off hyper-parameter. After this step, the target feature manifold will be closer to the source one while the classification accuracy of target samples is also preserved. We repeat the above process until it converges. The overall framework is shown in Fig. 2 and the training procedure is summarized in Algorithm 1.

4. Theoretical Analysis

In this section, we explain our motivation by briefly analyzing the relationship between our method and the theory of domain adaptation [1], which gives the upper bound of the expected error on the target domain as follows,

$$\forall h \in \mathcal{H}, R_{\mathcal{T}}(h) \leq R_{\mathcal{S}}(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T}) + \lambda, \quad (14)$$

where \mathcal{H} is the hypothesis class, $R_{\mathcal{S}}(h)$ is the expected error on the source domain which can be minimized explicitly since we have ground-truth source labels. $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ stands for the domain divergence, and λ is the error of the ideal joint hypothesis, i.e., $h^* = \arg \min_{h \in \mathcal{H}} R_{\mathcal{S}}(h) + R_{\mathcal{T}}(h)$. MCD employs two distinct hypotheses to reduce $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ and it treats λ as a negligible constant.

However, [4] shows there is an optimality gap between the optimal source hypothesis and the optimal target hypothesis in UDA, if two domains are misaligned at the category-level, we can hardly find a joint hypothesis that simultaneously minimizes the source and target expected errors, leading to a large λ . Hence, our motivation is to design a model that preserves the low $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{S}, \mathcal{T})$ in MCD and alleviates the problem caused by incorrect category-level alignment (i.e., minimizing λ). To this end, pseudo labels are necessary. Instead of using pseudo labels predicted by the source hypothesis as a direct supervision, which measures the loss of target samples *w.r.t.* the trained source hypothesis rather than the expected loss of the joint hypothesis

Algorithm 1 Cross-Domain Gradient Discrepancy Minimization for Unsupervised Domain Adaptation

Require: The set of labeled source samples $\{\mathcal{X}^s, \mathcal{Y}^s\}$ and the set of unlabeled target samples $\{\mathcal{X}^t\}$, the initialized generator G and classifiers F_1 and F_2 , the maximal epoch number N and the hyper-parameter α and β .

for epoch $\leftarrow 1$ to N **do**

step 1: Obtain the set of pseudo target labels $\{\mathcal{Y}^*\}$ through the Eq. (7) and Eq. (8). Then train G , F_1 and F_2 on both source and target samples:

$$\min_{\theta_g, \theta_{f_1}, \theta_{f_2}} \mathcal{L}_{cls}(\mathcal{X}^s, \mathcal{Y}^s) + \alpha \mathcal{L}_{cls}^W(\mathcal{X}^t, \mathcal{Y}^*),$$

step 2: Train F_1 and F_2 to maximize the divergence between the outputs of two classifiers on target samples without label information, as well as preserve the accuracy on source samples:

$$\min_{\theta_{f_1}, \theta_{f_2}} \mathcal{L}_{cls}(\mathcal{X}^s, \mathcal{Y}^s) - \mathcal{L}_{dis}(\mathcal{X}^t)$$

step 3: Train G to minimize the divergence between the outputs of two classifiers with gradient similarity constraint:

$$\min_{\theta_g} \mathcal{L}_{dis}(\mathcal{X}^t) + \beta \mathcal{L}_{GD}$$

end for

trained on them, we resort to minimizing the gradient discrepancy between domains. The insight of this trick is that we expect learning procedures of two domains not only finally yield a shared hypothesis, but also follow a similar optimization path. While this may restrict the optimization dynamics for the joint hypothesis, we argue that it enables a more fine-grained optimization and effective use of the incomplete optimizer, thus enabling a more accurate joint hypothesis and improving the accuracy of UDA.

5. Experiments

5.1. Dataset Description

DomainNet [33] is the largest and most challenging dataset to date for domain adaptation which contains about 600 thousand images distributed in 345 categories over 6 domains, including Clipart (clp), Infograph (inf), Painting (pnt), Quickdraw (qdr), Real (rel) and Sketch (skt). We adapt each domain to the other 5 domains.

VisDA-2017 [34] is a challenging large-scale dataset for UDA, which focuses on the simulation-to-reality shift. It consists of over 280K images across 12 categories. We use the training set and the validation set as the source domain and the target domain, respectively. The source domain contains 152,397 synthetic images generated by rendering 3D models and the target domain includes 55,388 real im-

Table 1. Accuracy(%) on **DomainNet** dataset for unsupervised domain adaptation (ResNet-50). We evaluate all pairwise transfers among 6 domains. The column-wise fields are applied as the source domain while the row-wise fields represent the target domain.

ResNet	clp	inf	pnt	qdr	rel	skt	Avg.	MCD	clp	inf	pnt	qdr	rel	skt	Avg.	BNM	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	14.2	29.6	9.5	43.8	34.3	26.3	clp	-	15.4	25.5	3.3	44.6	31.2	24.0	clp	-	12.1	33.1	6.2	50.8	40.2	28.5
inf	21.8	-	23.2	2.3	40.6	20.8	21.7	inf	24.1	-	24.0	1.6	35.2	19.7	20.9	inf	26.6	-	28.5	2.4	38.5	18.1	22.8
pnt	24.1	15.0	-	4.6	45.0	29.0	23.5	pnt	31.1	14.8	-	1.7	48.1	22.8	23.7	pnt	39.9	12.2	-	3.4	54.5	36.2	29.2
qdr	12.2	1.5	4.9	-	5.6	5.7	6.0	qdr	8.5	2.1	4.6	-	7.9	7.1	6.0	qdr	17.8	1.0	3.6	-	9.2	8.3	8.0
rel	32.1	17.0	36.7	3.6	-	26.2	23.1	rel	39.4	17.8	41.2	1.5	-	25.2	25.0	rel	48.6	13.2	49.7	3.6	-	33.9	29.8
skt	30.4	11.3	27.8	3.4	32.9	-	21.2	skt	37.3	12.6	27.2	4.1	34.5	-	23.1	skt	54.9	12.8	42.3	5.4	51.3	-	33.3
Avg.	24.1	11.8	24.4	4.7	33.6	23.2	20.3	Avg.	28.1	12.5	24.5	2.4	34.1	21.2	20.5	Avg.	37.6	10.3	31.4	4.2	40.9	27.3	25.3
CDAN	clp	inf	pnt	qdr	rel	skt	Avg.	SWD	clp	inf	pnt	qdr	rel	skt	Avg.	CGDM	clp	inf	pnt	qdr	rel	skt	Avg.
clp	-	13.5	28.3	9.3	43.8	30.2	25.0	clp	-	14.7	31.9	10.1	45.3	36.5	27.7	clp	-	16.9	35.3	10.8	53.5	36.9	30.7
inf	18.9	-	21.4	1.9	36.3	21.3	20.0	inf	22.9	-	24.2	2.5	33.2	21.3	20.0	inf	27.8	-	28.2	4.4	48.2	22.5	26.2
pnt	29.6	14.4	-	4.1	45.2	27.4	24.2	pnt	33.6	15.3	-	4.4	46.1	30.7	26.0	pnt	37.7	14.5	-	4.6	59.4	33.5	30.0
qdr	11.8	1.2	4.0	-	9.4	9.5	7.2	qdr	15.5	2.2	6.4	-	11.1	10.2	9.1	qdr	14.9	1.5	6.2	-	10.9	10.2	8.7
rel	36.4	18.3	40.9	3.4	-	24.6	24.7	rel	41.2	18.1	44.2	4.6	-	31.6	27.9	rel	49.4	20.8	47.2	4.8	-	38.2	32.0
skt	38.2	14.7	33.9	7.0	36.6	-	26.1	skt	44.2	15.2	37.3	10.3	44.7	-	30.3	skt	50.1	16.5	43.7	11.1	55.6	-	35.4
Avg.	27.0	12.4	25.7	5.1	34.3	22.6	21.2	Avg.	31.5	13.1	28.8	6.4	36.1	26.1	23.6	Avg.	36.0	14.0	32.1	7.1	45.5	28.3	27.2

ages cropped from the Microsoft COCO dataset [23].

ImageCLEF¹ is a standard dataset for ImageCLEF 2014 domain adaptation challenge. It is established by selecting 12 common categories shared by the following three datasets: Caltech-256 (C), ImageNet ILSVRC2012 (I) and PASCALVOC2012(P). There are 600 images in each domain and 50 for each category, which makes a good property for experiments.

5.2. Implementation Details

We choose PyTorch [31] framework for implementing our models. NVIDIA GeForce RTX 2080 Ti GPU is used as our hardware platform. Following MCD, we add the class balance loss in addition to the aforementioned framework to improve the diversity and accuracy in all experiments during the training procedure, the weight of which is fixed to 0.1 in this paper. For all experiments, we resize all images to $224 \times 224 \times 3$. The network architecture and the hyper-parameters setting are demonstrated as follows.

Network Architecture. In our experiments, we realize the generator G with the ResNet-50 (for DomainNet and ImageCLEF) or ResNet-101 (for VisDA-2017) [12] pre-trained on ImageNet [8] to extract features from raw images. Following [36, 14], the original fully connected (FC) layer is replaced with a bottleneck layer of 256 units and a three-layer FC network, which are employed as our classifier F_1 and F_2 , the unit number of the hidden layer is set to 1000 in all experiments. A dropout layer is utilized before each FC layer and a batch normalization (BN) layer is applied after that FC layer.

Hyper-parameters. We train the whole network in an end-to-end fashion through back-propagation. Momentum SGD algorithm is used to optimize the network parameters with momentum 0.9 and weight decay ratio $5e^{-4}$. The classifiers are trained with learning rate $1e^{-3}$ for DomainNet and ImageCLEF, $3e^{-4}$ for VisDA-2017 since it can easily converge. The learning of the pre-trained ResNet backbone is 10 times lower than that of the classifiers. The batch size is set to 32 and the trade-off parameters α and β are set to 0.1 and 0.01 respectively in all experiments. Note that we do not

¹<https://www.imageclef.org/2014/adaptation>

use data augmentation such as the ten-crop ensemble used in [25, 40] during the evaluation, for fair comparison, the experimental results cited from previous works are also the versions without data augmentation.

5.3. Results

Here we show the comparison between our CGDM framework and other well-known UDA baselines, especially the methods that are most related to our work (e.g. MCD and SWD), to verify that our formulation can significantly boost the accuracy *w.r.t.* to these baselines. Note that in our experiments, the classification results are obtained by averaging the outputs of two distinct classifiers. The quantitative results are described as follows.

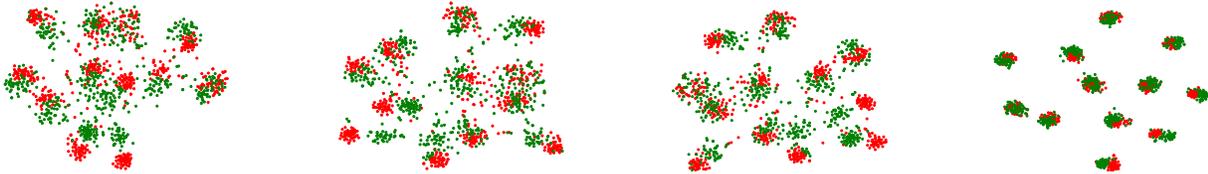
Results on DomainNet. In this experiment, we transfer each domain to the other five domains, and the class-wise average accuracy of each adaptation is reported in Table 1. From Table 1, our CGDM evidently suppresses other mainstream domain adaptation methods in mean accuracy.

The results are obtained after 10 epochs. In particular, both MCD and SWD perform significantly poorer than our method with 6.7% and 3.6% lower mean accuracy. The low accuracy of the source only model can infer that original target samples are likely to be distributed in a messy fashion such a challenging dataset with a large number of categories. Although the target samples outside the support of the source can be detected by the classifiers, they seem like to be matched to wrong decision boundaries, resulting in the wrong category-level alignment.

For other adversarial methods, they just forcibly align distributions and neglect the inherent gap between the source domain and the target domain, which may result in negative transfer. Our method applies a clustering-based self-supervised mechanism to improve the original accuracy of target samples, and the discrepancy between gradients of two domains is minimized to induce the generator to learn towards the direction of accurate classification for both source domain and target domain samples. The results on Table 1 suggest that our model can still achieve a more accurate adaptation in the target domain when facing such a large-scale dataset.

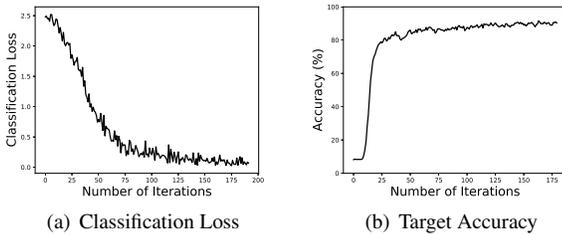
Table 2. Accuracy(%) on **VisDA-2017** dataset for unsupervised domain adaptation (ResNet-101).

Method	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg.
ResNet [12]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
DAN [24]	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.6
DANN [9]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
MinEnt [11]	80.3	75.5	75.8	48.3	77.9	27.3	69.7	40.2	46.5	46.6	79.3	16.0	57.0
MCD [36]	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
ADR [35]	87.8	79.5	83.7	65.3	92.3	61.8	88.9	73.2	87.8	60.0	85.5	32.3	74.8
SWD [14]	90.8	82.5	81.7	70.5	91.7	69.5	86.3	77.5	87.4	63.6	85.6	29.2	76.4
CDAN+E [25]	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.9
AFN [40]	93.6	61.3	84.1	70.6	94.1	79.0	91.8	79.6	89.9	55.6	89.0	24.4	76.1
BNM [7]	89.6	61.5	76.9	55.0	89.3	69.1	81.3	65.5	90.0	47.3	89.1	30.1	70.4
MSTN+DSBN [3]	94.7	86.7	76.0	72.0	95.2	75.1	87.9	81.3	91.1	68.9	88.3	45.5	80.2
CGDM (ours)	93.4	82.7	73.2	68.4	92.9	94.5	88.7	82.1	93.4	82.5	86.8	49.2	82.3



(a) ResNet-50 (b) DANN (c) MCD (d) Ours

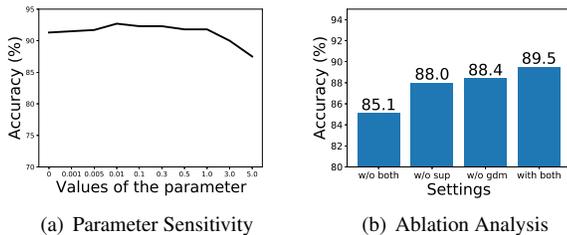
Figure 3. Visualization of features using t-SNE. We take C (red) \rightarrow I (green) on ImageCLEF as an example.



(a) Classification Loss

(b) Target Accuracy

Figure 4. Training process. We take C \rightarrow I on ImageCLEF as an example. The left figure depicts the classification loss during training. The right figure reports the accuracy of the target domain during the training.



(a) Parameter Sensitivity

(b) Ablation Analysis

Figure 5. Model analysis. The left figure shows the parameter sensitivity of our model (We take the task C \rightarrow I as an example). While the right figure shows the ablation study on ImageCLEF. The w/o is short for without, sup for self-supervised learning and gdm for gradient discrepancy minimization, respectively.

Results on VisDA-2017. We obtain the results on VisDA-2017 dataset after 10 epochs. The results in Table 2 illustrate our method can still outperform other popular approaches even if there is a large domain gap between syn-

thetic and real images. Specifically, our method performs much better than the source only model in all categories with the improvement up to 29.9% in terms of mean accuracy. For MCD and SWD which also apply the bi-classifier adversarial learning paradigm, our method also significantly suppresses them with the improvement of 10.4% and 5.9% respectively. In general, our method achieves the best in 5 categories: knife, person, plant, sktbrd and truck. In more difficult categories such as knife, sktbrd and truck, our method is much better than other existing methods, achieving 94.5%, 82.5% and 49.2% respectively. These excellent results strongly demonstrate the advantage of our method in improving the accuracy of the target domain.

Results on ImageCLEF. We access 6 types of adaptation scenarios on ImageCLEF and Table 3 reports the experimental results. We stop the training process after 100 epochs. The experimental results of the comparative methods are cited from previous papers, since some of them do not reported the randomness on this dataset, so we do not report their randomness too. Table 3 shows that our method outperforms other popular baselines and achieves the best average accuracy (89.5%). As for the difficult scenarios (e.g. C \rightarrow P), our method can obviously improve the accuracy of the target domain, which shows the effectiveness of our approach.

In addition, in scenarios that involve many categories (DomainNet) or there is a large domain gap (VisDA-2017), our model outperforms other methods by a large margin. We can empirically conclude that the the consideration of accuracy is particularly important in these scenarios.

Table 3. Accuracy(%) on **ImageCLEF** dataset for unsupervised domain adaptation (ResNet-50).

Method (Source \rightarrow Target)	I \rightarrow P	P \rightarrow I	I \rightarrow C	C \rightarrow I	C \rightarrow P	P \rightarrow C	Avg
ResNet [12]	74.8 \pm 0.3	83.9 \pm 0.1	91.5 \pm 0.3	78.0 \pm 0.2	65.5 \pm 0.3	91.2 \pm 0.3	80.7
DAN [24]	74.5 \pm 0.4	82.2 \pm 0.2	92.8 \pm 0.2	86.3 \pm 0.4	69.2 \pm 0.4	89.80 \pm 0.4	82.5
RTN [26]	75.6	86.8	95.3	86.9	72.7	92.2	84.9
DANN [9]	75.0 \pm 0.6	86.0 \pm 0.3	96.2 \pm 0.4	87.0 \pm 0.5	74.3 \pm 0.5	91.5 \pm 0.6	85.0
MinEnt [11]	76.2	85.7	93.5	83.5	69.3	89.7	83.0
MCD [36]	77.3	89.2	92.7	88.2	71.0	92.3	85.1
CDAN+E [25]	77.7 \pm 0.3	90.7 \pm 0.2	97.7\pm0.3	91.3 \pm 0.3	74.2 \pm 0.2	94.3 \pm 0.3	87.7
AFN [40]	79.3\pm0.1	93.3\pm0.4	96.3 \pm 0.4	91.7 \pm 0.0	77.6 \pm 0.1	95.3 \pm 0.1	88.9
BNM [7]	77.2	91.2	96.2	91.7	75.7	96.7	88.1
CGDM (ours)	78.7 \pm 0.2	93.3\pm0.1	97.5 \pm 0.3	92.7\pm0.2	79.2\pm0.1	95.7 \pm 0.2	89.5

5.4. Model Analysis

In this section, we take a further step to analyze the properties of the model in terms of convergence, feature visualization, parameter sensitivity and ablation study.

Feature Visualization. To further demonstrate the effectiveness of our model and have an intuitive understanding, we visualize the feature distribution learned by the model using t-SNE [29]. We take $C \rightarrow I$ on ImageCLEF as an example. It shows that in all scenarios, the source samples present a discriminative distribution. While in the absence of domain adaptation, the target samples likely to be disorganized, as shown in Fig. 3(a). Fig. 3(b) shows that using the domain adversarial neural network, the distribution discrepancy between the source domain and the target domain is reduced, while the discriminability of the target domain is relatively poor. Fig. 3(c) shows that MCD improves the discriminability of features. After using our method, the distribution of target samples aligns well with the source one at category-level. The result verifies the effectiveness and feasibility of our method.

Convergence. In order to verify the convergence tendency of our method, we report the classification loss on source samples in Fig. 4(a), this loss is obtained on the classifier F_1 . Fig. 4(b) depicts the accuracy curve of target samples during the training process. Here we take $C \rightarrow I$ on ImageCLEF as an example. These figures show that our method can markedly reduce the loss and improve the accuracy with the number of iterations increases, which proves that the training process is smooth and convergent.

Sensitivity to Hyper-Parameter. We check the hyper-parameter sensitivity of our model. Our framework involves two hyper-parameters α and β , the gradient discrepancy loss constitutes the main new methodological contribution in this paper while the self-supervised loss is used as an auxiliary loss, so we fix the α as 0.1 because we empirically find that it performs well in this value. Then we report the accuracy of $C \rightarrow I$ by turning β from 0 to 5, which is shown in Fig. 5(a). When β is between 0 and 1, there was no obvious deterioration in accuracy, when β is larger than 1, the accuracy begins to decline since the weight of the gradient discrepancy loss is larger than that of the supervised

learning loss on source samples. In general, our model is not sensitive to parameters.

Ablation Study. Our framework is composed of a self-supervised learning module and a gradient discrepancy minimization module. In order to test the effectiveness of each module, we conduct experiments without the self-supervised learning and the gradient discrepancy minimization, respectively. Limited by space, we take the mean accuracy on ImageCLEF as an example. As shown in Fig. 5(b), both two modules can significantly improve the accuracy compared to the vanilla bi-classifier adversarial learning, and the model performs better when they work together.

Conditional Gradient Discrepancy. We use two types of gradient discrepancy in this paper. In addition to align marginal gradients by regarding all samples in a batch as a whole, we have tried to align gradients for source and target samples within the same category separately. However, in our experiments, we could not see obvious improvement using the latter version. It indicates that the overall gradient of a batch is able to express the category information of the domain, thus we use the former one for less computation.

6. Conclusion

In this paper, we investigate the inaccurate issue of the conventional bi-classifier adversarial learning for domain adaptation. To alleviate this issue, we propose a novel UDA method which aims to minimize the gradient discrepancy between two domains to achieve a better distribution alignment at category-level. In addition, self-supervised learning is used to obtain more reliable pseudo labels of target samples. Extensive experimental results on large scale datasets demonstrate the advantage of our method.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 61806039, 62073059 and 61832001, and in part by Sichuan Science and Technology Program under Grant 2020YFG0080.

References

- [1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. **5**
- [2] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. **3, 4**
- [3] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019. **7**
- [4] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8869–8878, 2020. **5**
- [5] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1081–1090, 2019. **1**
- [6] Jaehoon Choi, Minki Jeong, Taekyung Kim, and Changick Kim. Pseudo-labeling curriculum for unsupervised domain adaptation. *arXiv preprint arXiv:1908.00262*, 2019. **2, 4**
- [7] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3941–3950, 2020. **7, 8**
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. **6**
- [9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189, 2015. **1, 2, 7, 8**
- [10] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. **1**
- [11] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005. **7, 8**
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. **6, 7, 8**
- [13] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902, 2019. **2, 3**
- [14] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019. **1, 2, 3, 6, 7**
- [15] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Zi Huang. Cycle-consistent conditional adversarial transfer networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 747–755, 2019. **2**
- [16] 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*, 2020. **2**
- [17] Jingjing Li, Mengmeng Jing, Ke Lu, Lei Zhu, and Heng Tao Shen. Locality preserving joint transfer for domain adaptation. *IEEE Transactions on Image Processing*, 28(12):6103–6115, 2019. **2**
- [18] Jingjing Li, Mengmeng Jing, Hongzu Su, Ke Lu, Lei Zhu, and Heng Tao Shen. Faster domain adaptation networks. *IEEE Transactions on Knowledge and Data Engineering*, 2021. **1**
- [19] Jingjing Li, Ke Lu, Zi Huang, Lei Zhu, and Heng Tao Shen. Heterogeneous domain adaptation through progressive alignment. *IEEE transactions on neural networks and learning systems*, 30(5):1381–1391, 2018. **1**
- [20] Jingjing Li, Ke Lu, Zi Huang, Lei Zhu, and Heng Tao Shen. Transfer independently together: A generalized framework for domain adaptation. *IEEE transactions on cybernetics*, 49(6):2144–2155, 2018. **2**
- [21] Shuang Li, Chi Harold Liu, Qiuxia Lin, Qi Wen, Limin Su, Gao Huang, and Zhengming Ding. Deep residual correction network for partial domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. **2**
- [22] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. *arXiv preprint arXiv:2002.08546*, 2020. **3, 4**
- [23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. **6**
- [24] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. **1, 2, 7, 8**
- [25] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018. **1, 2, 6, 7, 8**
- [26] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in neural information processing systems*, pages 136–144, 2016. **2, 8**

- [27] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217, 2017. 1, 2
- [28] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019. 2
- [29] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8
- [30] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010. 1
- [31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 6
- [32] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. *arXiv preprint arXiv:1809.02176*, 2018. 2
- [33] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 5
- [34] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 5
- [35] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017. 3, 7
- [36] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 1, 2, 3, 6, 7, 8
- [37] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 2
- [38] Rui Wang, Masao Utiyama, Lema Liu, Kehai Chen, and Eiichiro Sumita. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, 2017. 2
- [39] Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *International conference on machine learning*, pages 5423–5432. PMLR, 2018. 2, 4
- [40] 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 International Conference on Computer Vision*, pages 1426–1435, 2019. 6, 7, 8
- [41] Hongliang Yan, Zhetao Li, Qilong Wang, Peihua Li, Yong Xu, and Wangmeng Zuo. Weighted and class-specific maximum mean discrepancy for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 2019. 2
- [42] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*, 2017. 1, 2
- [43] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3801–3809, 2018. 2, 4
- [44] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5031–5040, 2019. 2
- [45] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2