Self-adaptive Re-weighted Adversarial Domain Adaptation

Shanshan Wang , Lei Zhang*
Learning Intelligence & Vision Essential (LiVE) Group
School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China
{wangshanshan, leizhang}@cqu.edu.cn

Abstract

Existing adversarial domain adaptation methods mainly consider the marginal distribution and these methods may lead to either under transfer or negative transfer. To address this problem, we present a self-adaptive re-weighted adversarial domain adaptation approach, which tries to enhance domain alignment from the perspective of conditional distribution. In order to promote positive transfer and combat negative transfer, we reduce the weight of the adversarial loss for aligned features while increasing the adversarial force for those poorly aligned measured by the conditional entropy. Additionally, triplet loss leveraging source samples and pseudo-labeled target samples is employed on the confusing domain. Such metric loss ensures the distance of the intra-class sample pairs closer than the inter-class pairs to achieve the class-level alignment. In this way, the high accurate pseudo-labeled target samples and semantic alignment can be captured simultaneously in the co-training process. Our method achieved low joint error of the ideal source and target hypothesis. The expected target error can then be upper bounded following Ben-David's theorem. Empirical evidence demonstrates that the proposed model outperforms state of the arts on standard domain adaptation datasets.

1 Introduction

Unsupervised Domain Adaptation (UDA) [Pan and Yang, 2010] task aims to recognize the unlabeled target domain data, leveraging a sufficiently labeled, related but different source domain. The key issue of UDA is to reduce distribution difference between the two domains, such that the learned classifier from source domain can well classify target domain samples. Generally, maximum mean discrepancy (MMD) [Long *et al.*, 2015], as a non-parametric metric, is commonly used to measure the dissimilarity of distributions. Recently, adversarial learning [Bousmalis *et al.*, 2016] has been successfully brought into UDA to reduce distribution discrepancy, in which domain-invariant or domain-confused feature representation is usually learned. Unlike many previous MMD-based methods, domain-adversarial neural networks focus on combining UDA and deep feature learning within a unified training paradigm. The goal of adversarial domain adaptation is to confuse the features between domains, so that domain-invariant representations are ultimately obtained.

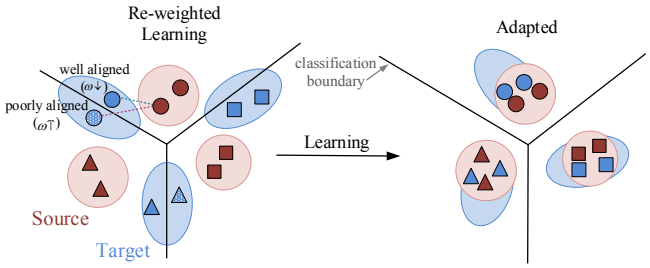

Figure 1: Motivation of our method. Let ω express weight, then down-weight ($\omega \downarrow$) well aligned samples and up-weight ($\omega \uparrow$) poorly aligned samples according to uncertainty. Different shapes represent different classes. Different colors mean different domains. Shadow denotes the misclassified samples.

However, as discussed in MCD [Saito *et al.*, 2018], it does not really guarantee safe domain alignment. *i.e.*, the alignment of category space between domains is ignored in alleviating domain shift. Target samples that are close to the decision boundary or far from their class centers could be misclassified by the classifier trained in source domain [Wen *et al.*, 2016]. Thus, previous domain adversarial adaptation methods that only match the domain distributions without exploiting the inner structures may be prone to under transfer (underfitting) or negative transfer (overfitting).

To address this problem, [Saito *et al.*, 2017] attempt to include the target samples into the learning of their models. Specifically, some method [Chen *et al.*, 2019] proposed to leverage pseudo-labels which is progressively guaranteed by an Easy-to-Hard strategy to learn target discriminative representations. These methods encourage a low-density separation between classes in the target domain.

However, as the domain bias exists, the pseudo labeled samples are not always correct. In order to obtain the high accurate labels of target samples, a much closer domain distribution is expected as the source classifier can generalize

*Contact Author

well on such domain-invariant target representations. In this paper, to tackle the aforementioned challenge, we take a two-step approach to learn domain invariant representations.

Firstly, the structure of adversarial network is adopted in our method. Although existing adversarial learning methods aim to reduce domain distribution discrepancy, they still suffer from a major limitation: these approaches mainly consider the marginal distribution while ignoring the conditional distribution, the classifier learned from source domain might be incapable of confidently distinguishing target samples. *i.e.*, the joint distributions of feature and category are not well aligned across domains. Thus, these methods may lead to either under transfer or negative transfer.

To promote positive transfer and combat negative transfer, we propose to recognize the transferability of each sample and re-weight these samples to force the underlying domain distributions closer. To push further along this line, we propose a self-adapted re-weighted adversarial DA approach shown in Fig. 1. Our method considers the transferable degree from the perspective of conditional distribution, therefore it can adapt better on target domain than previous approaches which only consider marginal distribution.

In information theory, the entropy is an uncertainty measure which can also be borrowed to quantify the adaptation. Different from previous methods which employ the conditional entropy directly, we utilize the entropy criterion to generate the weight and measure the degree of domain adaptation. Noteworthily, the conditional entropy is constructed by the conditional distribution. *i.e.*, our model does not just reduce conditional distributions between two domains directly, but dynamically leverages the conditional distribution to re-weight the samples. The inner reason lies that if the sample can get a high prediction by the conditional entropy, it can be regarded as a poorly-aligned sample, otherwise a well-aligned sample. The weights for those well aligned features are decreasing while for those poorly aligned features increasing in adversarial loss self-adaptively, then a better domain-level alignment can be achieved. If the distribution bias is reduced, the precisely pseudo-labeled samples in target domain can be chosen.

Secondly, as the pseudo labels are not always correct, they are not directly leveraged to train the classifier. Instead, they are employed to train the generalized feature representations. In our method, not only global domain-level aligning strategy, but also the metric loss is employed to learn the discriminative distance in the confusing domain. In our mechanism, triplet loss utilizes source samples and pseudo-labeled target samples to keep the samples align well in class-level. As a result, our model can learn better domain alignment features in the collaborative alignment adversarial training process and these features are not only domain invariant but also class discriminative for semantic alignment. This will have a more confident guarantee that the joint error of the ideal source and target hypothesis is low. The DA becomes possible as presented in Ben-David's theorem [Ben-David *et al.*, 2010].

The main contributions and novelties of this paper are summarized as follows.

- Our model attempts to learn the target generalized model to promote positive transfer and combat negative transfer. The network leveraging the joint distributions is much more proper than only the marginal distribution.
- In order to achieve the better domain confusion, we present a self-adaptive re-weighted adversarial domain adaptation approach through entropy from the perspective of conditional distribution.In our method, the adversarial network forces to reduce domain discrepancy by re-weighting the samples. The weights of the adversarial loss for well aligned features are decreased while increasing for those poorly aligned, such that a better domain-level alignment can be achieved.
- Besides the domain-level alignment, triplet loss is employed to enforce the features have better inter-class separation and intra-class compactness utilizing source samples and pseudo-labeled target samples. As a result, the feature representations which are not only domain invariant for domain alignment but also class discriminative for semantic alignment can be learned.

2 Related Work

Training CNN for UDA can be conducted through various strategies. Matching distributions of the middle features [Long *et al.*, 2015; Long *et al.*, 2017; Zellinger *et al.*, 2017] in CNN is considered to be effective for an accurate adaptation. These works pay attention to first-order or high-order statistics alignment.

Recent research on deep domain adaptation further embeds domain-adaptation modules in deep networks to boost transfer performance. In [Ganin *et al.*, 2016], DANN is proposed for domain adversarial learning, in which a gradient reversal layer is designed for confusing features from two domains. This method can be regarded as the baseline of adversarial learning methods. Tzeng *etal.* proposed an ADDA method [Tzeng *et al.*, 2017] which combines discriminative modeling, untied weight sharing and a GAN loss. Long *etal.* [Long *et al.*, 2018] also present a conditional adversarial domain adaptation (CDAN) that conditions the discriminative information conveyed into the predictions of classifier. Wang et al. [Wang *et al.*, 2019] propose a TADA focus the adaptation model on transferable regions or images both by local attention and global attention.

However, these methods are based on the theory that the predicted error is bounded by the distribution divergence. They do not consider the relationship between target samples and decision boundaries. To tackle these problems, multiple classifiers instead of the discriminator are considered and these methods become another branch. Saito et al. propose a ATDA method [Saito *et al.*, 2017] by tri-training three classifiers equally to give pseudo-labels to unlabeled samples. Later, he also proposes a new approach called MCD [Saito *et al.*, 2018] that uses two different classifiers to align those easily misclassified target samples through adversarial learning in CNN. Recently, Zhang et al. proposed SymNet [Zhang *et al.*, 2019] based on a symmetric design of source and target task classifiers, meanwhile, an additional classifier that shares with layer neurons was constructed.

In our method, we propose a different strategy to address these problems. Not only the source data but also the target

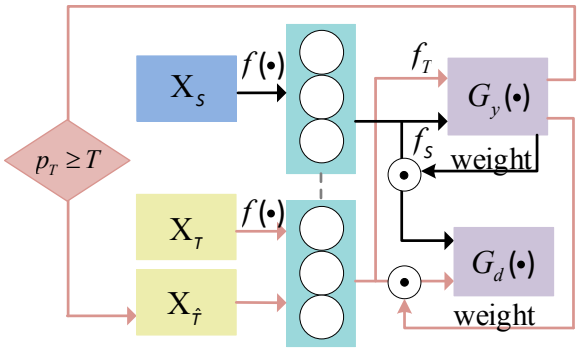

Figure 2: The framework of our method. To promote positive transfer and combat negative transfer, we utilize the entropy criterion to reveal the transferable degree of samples, then re-weight them and feed them into the discriminative network to force the underlying distributions closer.

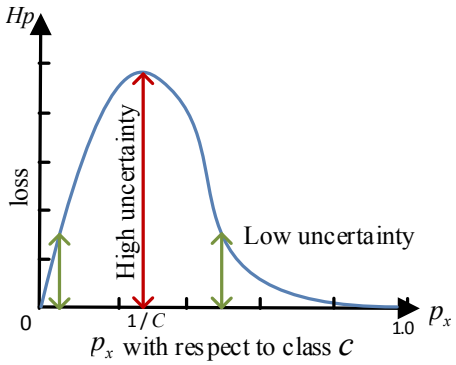

Figure 3: The conditional entropy measures the uncertainty. If the sample can get a low entropy, it can be regarded as a well-aligned transferable sample, else it is a poorly-aligned sample.

samples are leveraged to align domain features and class relations.

3 Self-adaptive Re-weighted Adversarial DA

An overview of our method is depicted in Fig. 2. In UDA, we suppose $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ and $\mathcal{D}_t = \{x_j^t\}_{j=1}^{n_t}$ to be the labeled source data and unlabeled target data, drawn from different distributions respectively. Our goal is to predict the target label $\hat{y}^t = \arg\max G_y(f(x^t))$ and minimize the target risk $\epsilon_t(G_y) = \mathbb{E}_{(x^t,y^t)\sim\mathcal{D}_t}[G_y(f(x^t)) \neq y^t]$, where $G_y(\cdot)$ represents the softmax output and $f(\cdot)$ refers to the feature representation.

Our method aims to construct a target generalized network. The re-weighted adversarial domain adaptation forces a close global domain-level alignment. Simultaneously, triplet loss is leveraged to train the class-discriminative representations utilizing source samples and pseudo-labeled samples. Then the classifier gradually increase accuracies on the target domain.

Preliminaries: Domain Adversarial Network. In DA setting, domain adversarial networks have been successfully explored to minimize the cross-domain discrepancy by extracting transferable features. The procedure is a two-player game: the first player is the domain discriminator G_d trained to distinguish the source domain from the target domain, and the second player is the feature extractor $f(\cdot)$ trained to confuse the domain discriminator. The objective function of domain adversarial network is as follows:

$$\begin{aligned} \mathcal{L}_{task}^s(\theta_f, \theta_y) &= \frac{1}{n_s} \sum_{\mathbf{x}_i \in \mathcal{D}_s} \mathcal{L}_y(G_y(f(\mathbf{x}_i)), \mathbf{y}_i^s), \\ \mathcal{L}_D(\theta_f, \theta_y, \theta_d) &= -\frac{1}{n_s + n_t} \sum_{\mathbf{x}_i \in (\mathcal{D}_s \cup \mathcal{D}_t)} \mathcal{L}_d(G_d(f(\mathbf{x}_i)), d_i), \end{aligned} \tag{1}$$

where θ_f, θ_d represent the parameters of feature network and domain discriminator, respectively. θ_y is the parameter of source classifier. d_i is the domain label of sample x_i.

Self-adaptive Re-weighted Adversarial DA. In practical domain adaptation problems, however, the data distributions of the source domain and target domain usually embody complex multimode structures. Thus, previous domain adversarial adaptation methods that only match the marginal distributions without exploiting the multimode structures maybe prone to either under transfer or negative transfer. To promote positive transfer and combat negative transfer, we should find a technology to reveal the transferable degree of samples and then re-weight them to force the underlying distribution closer.

As mentioned earlier, not all images are equally transferable in domain adaptation network and some images are more transferable than others. Therefore, we propose a method to measure adaptable degree using the certainty estimate. In information theory, the entropy functional is an uncertainty measure which nicely meets our need to quantify the adaptation and is depicted in Fig. 3. Therefore, we utilize the entropy criterion to estimate weights and improve the domain confusion by relaxing the alignment on these well aligned samples and focusing the alignment on these poorly aligned. If the sample can get a low entropy, it can be regarded as a well-aligned transferable sample, else it is a poorly aligned sample. The conditional distribution leveraged in the entropy is not considered in the standard adversarial DA methods. We adopt the conditional entropy as the indicator to weight the adversarial loss and the adversarial loss is extended as

$$\begin{aligned} \mathcal{L}_{adv}(\theta_f, \theta_d) &= -\frac{1}{n_s + n_t} \sum_{\mathbf{x}_i \in (\mathcal{D}_s \cup \mathcal{D}_t)} (1 + \mathcal{H}_p) \mathcal{L}_d(G_d(f(\mathbf{x}_i)), d_i), \\ where \quad \mathcal{H}_p &= -\frac{1}{C} \sum_{c=1}^{C} p_c \log(p_c) \end{aligned} \tag{2}$$

where C is the number of classes and p_c is the probability of predicting an sample to class c.

To make the best of conditional distribution, entropy minimization principle is adopted to enhance discrimination of learned models for target data by following [Long *et al.*, 2016]. In order to reduce wrong classified samples due to domain shift, the entropy minimization loss is used to update both the feature network and classifier.

$$\mathcal{L}_h(\theta_f, \theta_y) = -\frac{1}{n_t} \sum_{\mathbf{x}_i \in \mathcal{D}_t} p_t \log(p_t). \tag{3}$$

Class-level Alignment. So far, we only consider the global domain-level confusion, the discriminative power between classes is not involved [Hong *et al.*, 2015]. For one thing, samples with same labels should be pulled together in the embedding space. For another, samples with different labels

should be separated apart. Naturally, metric learning [Yang *et al.*, 2017] as an effective method can be implemented to achieve our goal. Triplet loss [Schroff *et al.*, 2015] tries to enforce a margin m between each pair of samples from one class to all other classes. It allows samples to enforce the distance and thus discriminate to other classes. In our paper, we select triplet loss to train the class-level aligned features.

Intuitively, a target generalized classifier is much more proper than the source classifier for the target domain. We propose to assign pseudo-labels to target samples and train the network as if they were true labels. Noteworthily, as the pseudo labels are not always correct, they are not directly leveraged to train the classifier. In order to make the best of pseudo-labeled samples, we leverage the source samples and pseudo-labeled target samples to construct the sample-pairs in metric learning. We follow the sampling strategy in [Deng *et al.*, 2018] to randomly select samples.

The pseudo label $\hat{y}_i^t$ of x_i^t is predicted based on maximum posterior probability using the source cross-entropy loss. It is progressively updated during optimization. Additionally, we only select target images with predicted scores above a high threshold T for building the semantic relations based on the intuitive consideration that the image with the high predicted score is more likely to be classified correctly. We empirically set the threshold T as a constant.

In the confusing domain, given an anchor image $\mathbf{x}_a$, a positive image $\mathbf{x}_p$, and a negative image $\mathbf{x}_n$, the minimized loss is as:

$$\mathcal{L}_{\text{tri}}(\theta_f) = \sum_{\substack{\mathbf{x}_i \in (\mathcal{D}_s \cup \mathcal{D}_t) \\ y_a = y_p \neq y_n}} [m + d_{a,p} - d_{a,n}]_+. \tag{4}$$

Overall Training Loss. With Eq. (1), Eq. (2), Eq. (3) and Eq. (4), the overall training loss of our model is given by,

$$\mathcal{L} = \mathcal{L}_{task}^s + \mathcal{L}_{adv} + \mathcal{L}_h + \mathcal{L}_{\text{tri}}. \tag{5}$$

The optimization problem is to find the parameters $\hat{\theta}_f, \hat{\theta}_y$ and $\hat{\theta}_d$ that jointly satisfy

$$\begin{aligned} (\hat{\theta}_f, \hat{\theta}_y) &= \arg\min_{\theta_f, \theta_y} \mathcal{L}(\theta_f, \theta_y, \theta_d) \\ (\hat{\theta}_d) &= \arg\max_{\theta_d} \mathcal{L}(\theta_f, \theta_y, \theta_d). \end{aligned} \tag{6}$$

4 Experiment

In this section, several benchmark datasets, not only the toy datasets as USPS+MNIST datasets, but also Office-31 dataset [Saenko *et al.*, 2010], ImageCLEF-DA [Long *et al.*, 2017] dataset and Office-Home [Venkateswara *et al.*, 2017] dataset, are adopted for evaluation.

Handwritten Digits Datasets. USPS (**U**) and MNIST (**M**) datasets are toy datasets for domain adaptation. They are standard digit recognition datasets containing handwritten digits from $0-9$. USPS consists of 7,291 training images and 2,007 test images of size 16×16. MNIST consists of 60,000 training images and 10,000 test images of size 28×28. We construct two tasks: $U \rightarrow M$ and $M \rightarrow U$ and we follow the experimental settings of [Hoffman *et al.*, 2018].

Office-31 Dataset. This dataset is a most popular benchmark dataset for cross-domain object recognition. The dataset consists of daily objects in an office environment and includes three domains such as Amazon (**A**), Webcam (**W**) and Dslr (**D**). There are 2,817 images in domain **A**, 795 images in **W** and 498 images in domain **D** making total 4,110 images. With each domain worked as source and target alternatively, 6 cross-domain tasks are formed, *e.g.*, $\mathbf{A} \rightarrow \mathbf{D}$ *etc*.

ImageCLEF-DA Dataset. The ImageCLEF-DA is a benchmark for ImageCLEF 2014 domain adaptation challenge. It contains 12 common categories shared by three public datasets: Caltech-256 (**C**), ImageNet ILSVRC 2012 (**I**) and Pascal VOC 2012 (**P**). In each domain, there are 50 images per class and totally 600 images are constructed. Images in ImageCLEF-DA are of equal size, making it a good alternative dataset. We evaluate all methods across three transfer domains and build 6 cross-domain tasks: *e.g.*, $\mathbf{I} \rightarrow \mathbf{P}$ *etc*.

Office-Home Dataset. This is a new and challenging dataset for domain adaptation, which consists of 15,500 images from 65 categories coming from four significantly different domains: Artistic images (**Ar**), Clip Art (**Cl**), Product images (**Pr**) and Real-World images (**Rw**). With each domain worked as source and target alternatively, there are 12 DA tasks on this dataset. The images of these domains have substantially different appearance and backgrounds, and the number of categories is much larger than that of Office-31 and ImageCLEF-DA, making it more difficult to transfer across domains.

Results. In our experiment, the target labels are unseen by following the standard evaluation protocol of UDA [Long *et al.*, 2017]. Our implementation is based on the PyTorch framework. For the toy datasets of handwritten digits, we utilize the LeNet. For other datasets, we use the pre-trained ResNet-50 as backbone network. We adopt the progressive training strategies as in CDAN [Long *et al.*, 2018]. In the process of selecting pseudo-labeled samples, the threshold T is empirically set as the constant 0.9. The margin m and N_0 in triplet loss are set as 0.3 and 3 following the setting as usual, respectively.

We evaluate the rank-1 classification accuracy for comparison. For handwritten digits, as there are plenty of different configutations in these datsets, in order to give the fair comparison, we only show some the recent results with the same backbone and training/test split. we compared with ADDA [Tzeng *et al.*, 2017], CoGAN [Liu and Tuzel, 2016], UNIT [Liu *et al.*, 2017], CYCADA [Hoffman *et al.*, 2018] and CDAN [Long *et al.*, 2018]. For other datasets, our compared baseline methods include DAN [Long *et al.*, 2015], DANN [Ganin *et al.*, 2016], JAN [Long *et al.*, 2017], CDAN [Long *et al.*, 2018]and SAFN [Xu *et al.*, 2019]. Besides, on Office-31 dataset, we compare with TCA [Pan *et al.*, 2011], GFK [Gong *et al.*, 2012], DDC [Tzeng *et al.*, 2014], RTN [Long *et al.*, 2016], ADDA [Tzeng *et al.*, 2017], MADA [Cao *et al.*, 2018], GTA [Sankaranarayanan *et al.*, 2019], MCD [Saito *et al.*, 2018], iCAN [Zhang *et al.*, 2018], TADA [Wang *et al.*, 2019] and SymNet [Zhang *et al.*, 2019]. On ImageCLEF-DA dataset, RTN [Long *et al.*, 2016], MADA [Cao *et al.*, 2018] and iCAN [Zhang *et al.*, 2018] are compared. On Office-Home dataset, TADA [Wang *et al.*, 2019] and SymNet [Zhang *et al.*, 2019] are compared.

OfficeHome	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw →Pr	Avg.
ResNet-50	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DAN	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
JAN	45.9	61.2	68.9	50.4	59.7	61.0	45.8	43.4	70.3	63.9	52.4	76.8	58.3
CDAN	50.7	70.6	76.0	57.6	70.0	70.0	57.4	50.9	77.3	70.9	56.7	81.6	65.8
SAFN	52.0	71.7	76.3	**64.2**	69.9	71.9	63.7	51.4	77.1	70.9	57.1	81.5	67.3
TADA	53.1	72.3	77.2	59.1	71.2	72.1	59.7	53.1	78.4	72.4	60.0	82.9	67.6
SymNet	47.7	72.9	78.5	**64.2**	71.3	**74.2**	**64.2**	48.8	79.5	**74.5**	52.6	82.7	67.6
Ours	**55.5**	**73.5**	**78.7**	60.7	**74.1**	73.1	59.5	**55.0**	**80.4**	72.4	**60.3**	**84.3**	**68.9**

Table 4: Recognition accuracy (%) on Office-Home dataset. All models utilize ResNet-50 as base architecture.

Handwritten	M → U	U → M	Avg.
ADDA	89.4	90.1	89.8
CoGAN	95.6	93.1	94.3
UNIT	**96.0**	93.6	94.8
CDAN	93.9	96.9	95.4
CYCADA	95.6	96.5	**96.1**
Ours	94.1	**98.0**	**96.1**

Table 1: Recognition accuracies (%) on handwritten digits datasets. All models utilize LeNet as base architecture.

Office-31	A→W	D→W	W→D	A→D	D→A	W→A	Avg.
Source Only	68.4	96.7	99.3	68.9	62.5	60.7	76.1
TCA	72.7	96.7	99.6	74.1	61.7	60.9	77.6
GFK	72.8	95.0	98.2	74.5	63.4	61.0	77.5
DDC	75.6	96.0	98.2	76.5	62.2	61.5	78.3
DAN	80.5	97.1	99.6	78.6	63.6	62.8	80.4
RTN	84.5	96.8	99.4	77.5	66.2	64.8	81.6
DANN	82.0	96.9	99.1	79.7	68.2	67.4	82.2
ADDA	86.2	96.2	98.4	77.8	69.5	68.9	82.9
JAN	85.4	97.4	99.8	84.7	68.6	70.0	84.3
MADA	90.0	97.4	99.6	87.8	70.3	66.4	85.2
SAFN	88.8	98.4	99.8	87.7	69.8	69.7	85.7
GTA	89.5	97.9	99.8	87.7	72.8	71.4	86.5
MCD	88.6	98.5	**100.0**	92.2	69.5	69.7	86.5
iCAN	92.5	**98.8**	**100.0**	90.1	72.1	69.9	87.2
CDAN	94.1	98.6	**100.0**	92.9	71.0	69.3	87.7
TADA	94.3	98.7	99.8	91.6	72.9	73.0	88.4
SymNet	90.8	**98.8**	**100.0**	**93.9**	**74.6**	72.5	88.4
Ours	**95.2**	98.6	**100.0**	91.7	74.5	**73.7**	**89.0**

Table 2: Recognition accuracies (%) on the Office31 dataset. All models utilize ResNet-50 as base architecture.

5 Discussion

Ablation Study. The ablation analysis results under different model variants with some loss removed are presented in Table 5. The baseline of *ResNet-50* denotes that only the source classifier based on cross-entropy loss is trained. *DANN* is another baseline, in which the cross-entropy loss and domain alignment are taken into account, and the performance is increased from 74.3% to 82.1%. Besides, we additionally optimize the entropy minimization loss of target samples over their feature extractors and denote them as *ResNet-50 (Em)* and *DANN (Em)*, respectively. They two can be regarded as another baselines compared with our method.

Our model benefits from both the novel re-weighted domain adaptation and class-level alignment cross domains. To

ImageCLEF-DA	I→P	P→I	I→C	C→I	C→P	P→C	Avg.
Source Only	74.8	83.9	91.5	78.0	65.5	91.2	80.7
DAN	74.5	82.2	92.8	86.3	69.2	89.8	82.5
RTN	75.6	86.8	95.3	86.9	72.7	92.2	84.9
DANN	75.0	86.0	96.2	87.0	74.3	91.5	85.0
JAN	76.8	88.0	94.7	89.5	74.2	91.7	85.8
MADA	75.0	87.9	96.0	88.8	75.2	92.2	85.8
iCAN	**79.5**	89.7	94.7	89.9	**78.5**	92.0	87.4
CDAN	77.7	90.7	**97.7**	**91.3**	74.2	94.3	87.7
SAFN	78.0	91.7	96.2	91.1	77.0	94.7	88.1
Ours	78.3	**91.3**	96.7	90.5	78.1	**96.2**	**88.5**

Table 3: Recognition accuracies (%) on ImageCLEF-DA. All models utilize ResNet-50 as base architecture.

Office-31	A → W	W → D	A → D	W → A	Avg.
ResNet-50	68.4	99.3	68.9	60.7	74.3
ResNet-50 (Em)	89.3	**100.0**	89.2	69.0	86.9
DANN	82.0	99.1	79.7	67.4	82.1
DANN (Em)	89.8	**100.0**	90.1	69.0	87.2
DANN (Em+$\mathcal{H}_p$)	92.3	100.0	91.1	71.9	88.8
DANN (Em+$\mathcal{L}_{tri-s}$)	92.4	99.7	90.3	71.9	88.6
DANN (Em+$\mathcal{L}_{tri}$)	93.8	99.8	**91.7**	72.6	89.5
Ours	**95.2**	**100.0**	**91.7**	**73.7**	**90.2**

Table 5: Ablation study on the Office-31 dataset.

investigate how different components in our model, we add every item alternately. We can verify the items from two aspects. On the one hand, we do not re-weight the adversarial loss and the *DANN (Em)* can be regarded as the baseline. Then we add the cross-domain weight $\mathcal{H}_p$ into the adversarial loss, the training setting of which is denoted as "*DANN (Em+$\mathcal{H}_p$)*", the performance is increased from 87.2% to 88.8% after adding re-weight item $\mathcal{H}_p$. On the other, we add the metric triplet loss $\mathcal{L}_{tri}$ into the *DANN* baseline, the training setting of which is denoted as "*DANN (Em +$\mathcal{L}_{tri}$)*" and the performance is increased from 87.2% to 89.5%. Furthermore, the item of "*DANN (Em +$\mathcal{L}_{tri}$)*" can be regarded as another baseline compared with "*Ours*" to prove the effectiveness of re-weight item $\mathcal{H}_p$ and the performance is increased to 90.2%. Additionally, because we also take some pseudo target labels into consideration, so for validating the triple loss with target samples, we have experimentally demonstrated the effectiveness of pseudo labels in metric loss. The performance is decreased from 89.5% to 88.6% after removing the target pseudo labels, *i.e.*, "*DANN (Em +$\mathcal{L}_{tri-s}$)*".

Quantitative Distribution Discrepancy. $\mathcal{A}$-distance

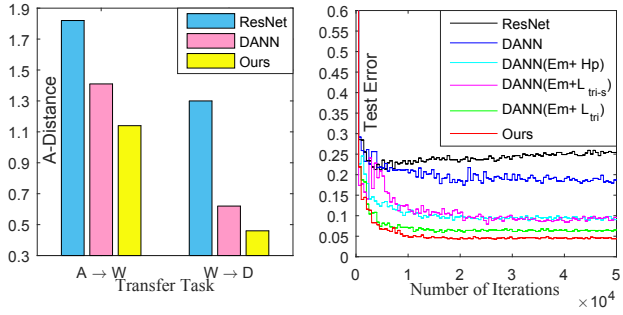

Figure 4: Illustration of model analysis: (a) Quantitative distribution discrepancy measured by $\mathcal{A}$-distance after domain adaptation. (b) Convergence on the test errors of different models.

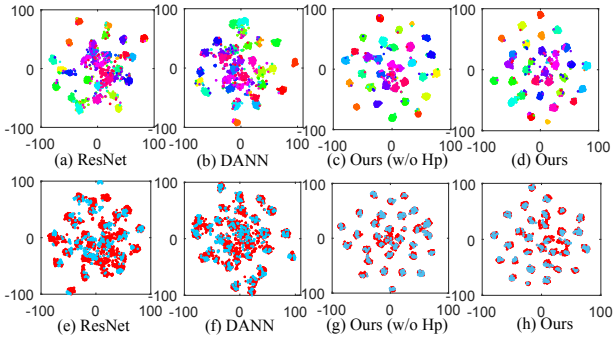

Figure 5: Feature visualization with t-SNE algorithm.

[Ben-David *et al.*, 2010] which jointly formulates source and target risk, are used to measure the distribution discrepancy after domain adaptation. $\mathcal{A}$-distance is defined as $d_{\mathcal{A}} = 2(1-2\epsilon)$, where ϵ is the classification error of a binary domain classifier (*e.g.*, SVM) for discriminating the source and target domains. Therefore, with the increasing discrepancy between two domains, the error ϵ becomes smaller. Figure 4 (a) shows $\mathcal{A}$-distance on different tasks by using different models. From Figure 4 (a), it is obvious that a large $\mathcal{A}$-distance denotes a large domain discrepancy. The distribution discrepancy analysis based on $\mathcal{A}$-distance in Office-31 dataset on tasks $\mathbf{A} \rightarrow \mathbf{W}$ and $\mathbf{W} \rightarrow \mathbf{D}$ is conducted by using *ResNet*, *DANN* and our complete model, respectively.

We can observe that $\mathcal{A}$-distance between domains after using our model is smaller than that of other two baselines, which suggests that our model is more effective in reducing the domain discrepancy gap. By comparing the distribution discrepancy between $\mathbf{A} \rightarrow \mathbf{W}$ and $\mathbf{W} \rightarrow \mathbf{D}$, obviously, $\mathbf{W} \rightarrow \mathbf{D}$ has a much smaller $\mathcal{A}$-distance than $\mathbf{A} \rightarrow \mathbf{W}$. From the classification accuracy in Table 2, the recognition rate of $\mathbf{W} \rightarrow \mathbf{D}$ is 100.0%, which is higher than $\mathbf{A} \rightarrow \mathbf{W}$ (95.2%). Therefore, the reliability of $\mathcal{A}$-distance is demonstrated.

Convergence. Figure 4 (b) shows the convergence of *ResNet*, *DANN*, our baseline method only with re-weight item, *i.e.*, *DANN (Em+*$\mathcal{H}_p$*)*, our baseline method with only source labels for triplet loss, *i.e.*, *DANN (Em+*$\mathcal{L}_{tri-s}$*)*, our baseline method with only triplet loss, *i.e.*, *DANN (Em+*$\mathcal{L}_{tri}$*)*, and our complete model *Ours*, respectively. We choose the task $\mathbf{A} \rightarrow \mathbf{W}$ in Office-31 dataset as an example and the test errors of different methods with the increasing number of iterations are shown in Figure 4 (b).

Feature Visualization. We visualize the domain invariant features learned by *ResNet*, *DANN*, *Ours* (w/o $\mathcal{H}_p$), and our complete model for further validating the effectiveness. For feature visualization, t-SNE visualization method is employed on the source domain and target domain in the $\mathbf{A} \rightarrow \mathbf{W}$ task from Office-31 dataset. The results of feature visualization for *ResNet* (traditional CNN), *DANN* (with adversarial learning), *Ours* (w/o $\mathcal{H}_p$) (*i.e.*, our model without re-weight), and our complete model are illustrated in Figure 5.

Note that Figure 5 (a)-(d) represent the results of source features from 31 classes marked in different colors, from which we observe that *Ours* (w/o $\mathcal{H}_p$) and our model can reserve better discrimination than other two baselines as the two consider the discriminative power. The features of two domains are visualized in Figure 5 (e)-(h). It is obvious that the features learned by *ResNet* across source and target domains can not be well aligned, without considering the feature distribution discrepancy. In *DANN*, by aligning the domain distribution, the distribution discrepancy of learned features between two domains can be improved. However, the class discrepancy of features from *DANN* is not improved, as *DANN* does not take the class level distribution into account. In our method, compared with *Ours* (w/o $\mathcal{H}_p$), it can alleviate domain discrepancy to some extent by our re-weight. From the classification accuracies in Table 5, *Ours* (95.2%) is a little better than *Ours* (w/o $\mathcal{H}_p$) (93.8%). From the results, the features learned by our model can be well aligned between two domains, but reserve more class discrimination including intra-class compactness and inter-class separability.

6 Conclusion

To promote positive transfer and combat negative transfer in DA problem, we propose a self-adaptive re-weighted adversarial approach that tries to enhance domain alignment from the perspective of conditional distribution. For alleviating the domain bias issue, on one hand, considering that not all images are equally transferable in domain adaptation network and some images are more transferable than others, we propose a method to reduce domain-level discrepancy by re-weighting the transferable samples. Our method reduces the weights of the adversarial loss for aligned features while increasing the adversarial forces for those poorly aligned adaptively. On the other, triplet loss is employed on the confusing domain to ensure the distance of the intra-class sample pairs closer than the inter-class pairs to achieve class-level alignment. Therefore, the high accurate pseudo-labeled target samples and semantic alignment can be captured simultaneously in this co-training process. The experimental results verify that the proposed model outperforms state-of-the-arts in various UDA tasks.

Acknowledgements

This work was supported by the National Science Fund of China under Grants (61771079), Chongqing Youth Talent Program, and the Fundamental Research Funds of Chongqing (No. cstc2018jcyjAX0250).

References

[Ben-David *et al.*, 2010] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. volume 79, 2010.

[Bousmalis *et al.*, 2016] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *NIPS*, 2016.

[Cao *et al.*, 2018] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *AAAI*, 2018.

[Chen *et al.*, 2019] Chaoqi Chen, Weiping Xie, Tingyang Xu, Wenbing Huang, Yu Rong, and Ding Xinghao. Progressive feature alignment for unsupervised domain adaptation. In *CVPR*, 2019.

[Deng *et al.*, 2018] Weijian Deng, Liang Zheng, and Jianbin Jiao. Domain alignment with triplets. *arXiv preprint arXiv:1812.00893*, 2018.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Franois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.

[Gong *et al.*, 2012] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073, 2012.

[Hoffman *et al.*, 2018] Judy Hoffman, Eric Tzeng, Taesung Park, Jun Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018.

[Hong *et al.*, 2015] Richang Hong, Yang Yang, Meng Wang, and Xian-Sheng Hua. Learning visual semantic relationships for efficient visual retrieval. *IEEE Trans on Big Data*, 1(4), 2015.

[Liu and Tuzel, 2016] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016.

[Liu *et al.*, 2017] Ming Yu Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *NIPS*, 2017.

[Long *et al.*, 2015] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015.

[Long *et al.*, 2016] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *NIPS*, 2016.

[Long *et al.*, 2017] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *ICML*, 2017.

[Long *et al.*, 2018] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *NIPS*, 2018.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 22(10), 2010.

[Pan *et al.*, 2011] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Networks*, 22(2):199–210, 2011.

[Saenko *et al.*, 2010] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. *ECCV*, 2010.

[Saito *et al.*, 2017] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *ICML*, 2017.

[Saito *et al.*, 2018] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. *CVPR*, 3, 2018.

[Sankaranarayanan *et al.*, 2019] Swami Sankaranarayanan, Yogesh Balaji, Carlos D. Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *CVPR*, 2019.

[Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[Tzeng *et al.*, 2014] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv*, 2014.

[Tzeng *et al.*, 2017] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. 2017. CVPR.

[Venkateswara *et al.*, 2017] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017.

[Wang *et al.*, 2019] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *AAAI*, 2019.

[Wen *et al.*, 2016] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

[Xu *et al.*, 2019] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *ICCV*, 2019.

[Yang *et al.*, 2017] Xun Yang, Meng Wang, and Dacheng Tao. Person re-identification with metric learning using privileged information. *IEEE Trans TIP*, 27(2), 2017.

[Zellinger *et al.*, 2017] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *ICLR*, 2017.

[Zhang *et al.*, 2018] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *CVPR*, 2018.

[Zhang *et al.*, 2019] Yabin Zhang, Hui Tang, Kui Jia, and Mingkui Tan. Domain-symmetric networks for adversarial domain adaptation. In *CVPR*, 2019.