Learning to Detect Open Classes for Universal Domain Adaptation

Bo Fu *, Zhangjie Cao *, Mingsheng Long (✉), and Jianmin Wang

School of Software, BNRist, Tsinghua University, China
Research Center for Big Data, Tsinghua University, China
{microhhh9,caozhangjie14}@gmail.com, {mingsheng,jimwang}@tsinghua.edu.cn

Abstract. Universal domain adaptation (UniDA) transfers knowledge between domains without any constraint on the label sets, extending the applicability of domain adaptation in the wild. In UniDA, both the source and target label sets may hold individual labels not shared by the other domain. A *de facto* challenge of UniDA is to classify the target examples in the shared classes against the domain shift. A more prominent challenge of UniDA is to mark the target examples in the target-individual label set (open classes) as "unknown". These two entangled challenges make UniDA a highly under-explored problem. Previous work on UniDA focuses on the classification of data in the shared classes and uses per-class accuracy as the evaluation metric, which is badly biased to the accuracy of shared classes. However, accurately detecting open classes is the mission-critical task to enable real universal domain adaptation. It further turns UniDA problem into a well-established close-set domain adaptation problem. Towards accurate open class detection, we propose Calibrated Multiple Uncertainties (CMU) with a novel transferability measure estimated by a mixture of uncertainty quantities in complementation: entropy, confidence and consistency, defined on conditional probabilities calibrated by a multi-classifier ensemble model. The new transferability measure accurately quantifies the inclination of a target example to the open classes. We also propose a novel evaluation metric called H-score, which emphasizes the importance of both accuracies of the shared classes and the "unknown" class. Empirical results under the UniDA setting show that CMU outperforms the state-of-the-art domain adaptation methods on all the evaluation metrics, especially by a large margin on the H-score.

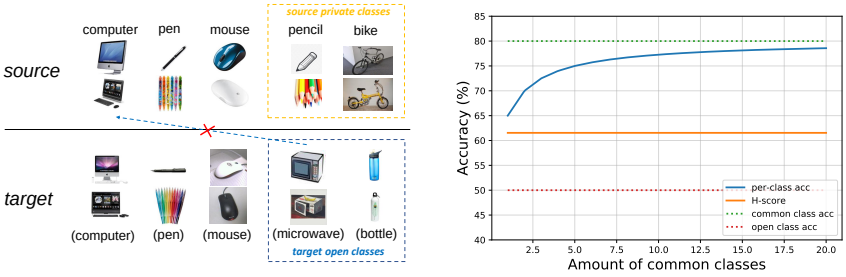
Keywords: Universal Domain Adaptation, Open Class Detection

1 Introduction

Domain adaptation (DA) relieves the requirement of labeled data in deep learning by leveraging the labeled data from a related domain [28]. Most DA methods constrain the source and target label sets to some extent, which are easily violated in complicated practical scenarios. For example, we can access molecule datasets

*Equal contribution

with annotated properties [39]. However, when predicting unknown molecules, we are exposed to two challenges: **(1)** The molecule structures such as scaffolds [13] may vary between training and testing sets, causing large *domain shift*; **(2)** Some molecules have property values never existing in our dataset such as unknown toxicity, which causes the *category shift*. To address the challenges, Universal Domain Adaptation (UniDA) [41] is raised to remove all label set constraints.



(a) The UniDA Setting (b) Metric

Fig. 1. (a) The UniDA Setting. There are 3 common, 2 source private and 2 target private classes. The red cross means that the open class "microwave" is easily misclassified to "computer". (b) Comparison of per-class accuracy and H-score. Assuming that the amount of samples in each category is equal. The classification accuracy of common classes is 80%, and the accuracy of open classes is 50%.

As shown in Figure 1(a), in UniDA, given any labeled source domain and unlabeled target domain, we need to classify target data correctly if it belongs to the common label set or mark it as "unknown" otherwise. UniDA poses two technical challenges: (**1**) Distribution matching is still needed but should be constrained into the common label set; (**2**) As a new challenge, we need to detect data of the target open classes without any target labeled data or prior knowledge. Detecting open classes is the key to UniDA since it can directly solve the second challenge, and if it is solved, the first challenge can be easily addressed by remove the open class data and perform partial domain adaptation methods.

Universal Adaptation Network (UAN) [41] addresses the challenges by quantifying the transferability of each sample based on the uncertainty and the domain similarity. However, as we analyzed in Section 3.1, the transferability suffers from two shortcomings. First, they use entropy to measure uncertainty and auxiliary domain classifier to measure domain similarity. Entropy lacks discriminability for uncertain and sharp predictions, especially with a large number of classes. The predictions of the auxiliary domain classifier are mostly overconfident as shown in Figure 4(b) in [41]. Second, the uncalibrated predictions make the transferability unreliable. Thus, UAN cannot detect open classes clearly. Such failure is hidden by the per-class accuracy used by UAN [41], which, as shown in Fig. 1(b), overly

focuses on the common label set, especially under large-scale classes. How to detect open classes and how to evaluate UniDA are still unsolved problems.

In this paper, we propose **Calibrated Multiple Uncertainties (CMU)** with a novel measurement to quantify the transferability of each sample. We improve the quality of the transferability over the previous work in two aspects. 1) We design a new uncertainty measurement by compensating entropy with consistency and confidence for the lack of ability to tackle particular predictions; 2) The multi-classifier architecture for uncertainty computation naturally forms an ensemble, which is the most suitable calibration method for the domain shift setting. The new transferability can more accurately estimate the uncertainty and more clearly differentiate different samples by uncertainty, which improves the accuracy of open class detection. Furthermore, we propose a new evaluation metric called H-score as the harmonic mean of the accuracy on common label set and the accuracy of marking data in the target private label set as "unknown". As shown in Fig. 1(b), the new criterion is high only when target data in both common and private label sets are classified accurately.

The main contributions of this paper are:

(1) We emphasize the importance of detecting open classes for UniDA. We propose Calibrated Multiple Uncertainties (CMU) with a novel transferability composed of entropy, consistency, and confidence. The three uncertainties are complementary to discriminate different degrees of uncertainty clearly and are well-calibrated by multiple classifiers, which distinguish target samples from common classes and open classes more clearly.

(2) We point out that the evaluation metric: per-class accuracy, used by UAN highly biases to common classes but fails to test the ability to detect open classes, especially when the number of common classes is large. We design a new evaluation protocol: H-score, as the harmonic mean of target common data accuracy and private data accuracy. It evaluates a balance ability to classify common class samples and filter open class samples.

(3) We conduct experiments on UniDA benchmarks. Empirical results show that CMU outperforms UAN and methods of other DA settings on all evaluation metrics, especially on the H-score. Deeper analyses show that the proposed transferability can distinguish the common label set from the open classes effectively.

2 Related Work

Domain adaptation settings can be divided into closed set, partial, open set domain adaptation and universal domain adaptation based on the label set relationship. Universal domain adaptation removes all constraints on the label set and includes all other domain adaptation settings.

Closed Set Domain Adaptation assumes both domains share the same label set. Early deep closed set domain adaptation methods minimize Maximum Mean Discrepancy (MMD) on deep features [34,20,22]. Recently, methods based on adversarial learning [8,33,21] are proposed to play a two-player minimax game between the feature extractor and a domain discriminator. Adversarial learning

methods achieves the state-of-the-art performance, which is further improved by recent works [29,15,40,31,24,19,37,12,45,23,5,14] with new architecture designs.

Partial Domain Adaptation requires that the source label set contains the target label set [2,44,3,4,11], which receives much more attention with access to large annotated dataset such as ImageNet [6] and Open Image [36]. To solve partial domain adaptation, one stream of works [2,3] uses target prediction to construct instance- and class-level weight to down-weight source private samples. Another stream [44,4] employs an auxiliary domain discriminator to quantify the domain similarity. Recent work [11] integrates the two weighting mechanisms.

Open Set Domain Adaptation (OSDA) is proposed by Busto *et al.* [25] to have private and shared classes in both domains but know shared labels. They use an Assign-and-Transform-Iteratively (ATI) algorithm to address the problem. Lian *et al.* [17] improves it by using entropy weight. Saito *et al.* [30] relaxed the problem by requiring no source private labels, so the target label set contains the source. Later OSDA methods [43,18,1] follow this more challenging setting and attack it by image translation [43] or a coarse-to-fine filtering process [18].

However, closed set, partial, open set domain adaptation are all restricted by label set assumptions. The latter two shed light on practical domain adaptation.

Universal Domain Adaptation (UniDA) [41] is the most general setting of domain adaptation, which removes all constraints and includes all the previous adaptation settings. It introduces new challenges to detect open classes in target data even with private classes in the source domain. UAN [41] evaluates the transferability of examples based on uncertainty and domain similarity. However, the uncertainty and domain similarity measurements, which are defined as prediction entropy and output of the auxiliary domain classifier, are not robust and discriminable enough. We propose a new uncertainty measurement as the mixture of entropy, consistency and confidence and design a deep ensemble model to calibrate the uncertainty, which characterizes different degrees of uncertainty and distinguishes target data in common label set from those in private label set.

3 Calibrated Multiple Uncertainties

In Universal Domain Adaptation (UniDA), a labeled source domain $\mathcal{D}^s = \{(\mathbf{x}^s, \mathbf{y}^s)\}$ and a unlabeled target domain $\mathcal{D}^t = \{(\mathbf{x}^t)\}$ are provided at training. Note that the source and target data are sampled from different distributions p and q respectively. We use $\mathcal{C}^s$ and $\mathcal{C}^t$ to denote the label set of the source domain and the target domain. $\mathcal{C} = \mathcal{C}^s \cap \mathcal{C}^t$ is the common label set shared by both domains while $\overline{\mathcal{C}}^s = \mathcal{C}^s \setminus \mathcal{C}$ and $\overline{\mathcal{C}}^t = \mathcal{C}^t \setminus \mathcal{C}$ are the label sets private to source and target respectively. $p_{\mathcal{C}^s}$ and $p_{\mathcal{C}}$ are used to denote the distributions of source data with labels in the label set $\mathcal{C}^s$ and $\mathcal{C}$ respectively, and $q_{\mathcal{C}^t}$, $q_{\mathcal{C}}$ are defined similarly. Note that the target label set is not accessible at training and only used for defining the UniDA problem. UniDA requires a model to distinguish target data in $\mathcal{C}$ from those in $\overline{\mathcal{C}}^t$, as well as predict accurate label for target data in $\mathcal{C}$.

3.1 Limitations of Previous Works

The most important challenge for UniDA is detecting open classes. We compare several state-of-the-art domain adaptation methods with open class detection module in Table 1 including UniDA method, UAN [41], and open set DA methods, STA [18] and OSBP [30]. STA and OSBP both use the confidence for an extra class as the criterion to detect open classes. However, as stated below, confidence alone lacks discriminability for particular predictions. In UAN, transferability is derived from uncertainty and domain similarity. Optimally, uncertainty is a well-established measurement to distinguish samples from $\mathcal{C}$ and from $\overline{\mathcal{C}}^s$ and $\overline{\mathcal{C}}^t$. But the uncertainty is measured by entropy, which lacks discriminability for uncertain and extremely sharp predictions. For the domain similarity, the auxiliary domain classifier is trained with domain label by supervised learning. So the predictions are over-confident. All the open class detection criteria before are unilateral and lack the discriminability for particular predictions.

Furthermore, the confidence for STA and OSBP and the uncertainty and domain similarity for UAN are based on uncalibrated prediction, meaning the prediction does not reflect the exact confidence, uncertainty or domain similarity of the sample. So all the criteria before are not estimated accurately and thus fail to distinguish target data in the common label set from the private label set.

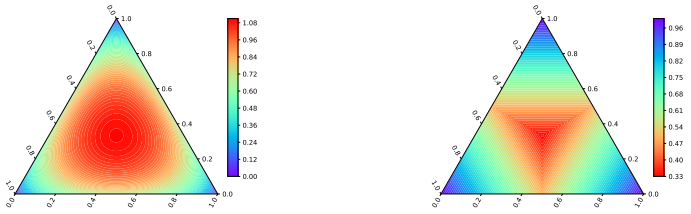
Table 1. Comparison of open class detection criterion for different methods

Criterion	Calibration	Entropy	Confidence	Consistency	Domain Similarity
OSBP [30]	✗	✗	✓	✗	✗
STA [18]	✗	✗	✓	✗	✗
UAN [41]	✗	✓	✗	✗	✓
CMU	✓	✓	✓	✓	✗

3.2 Multiple Uncertainties

We design a novel transferability to detect open class. We adopt the assumption made by UAN: the target data in $\mathcal{C}$ have lower uncertainty than target data in $\bar{\mathcal{C}}^t$. A well-defined uncertainty measurement should distinguish different degrees of uncertainty, e.g., distinguishing definitely uncertain predictions from slightly uncertain ones. Then we can rank the uncertainty of target samples and mark the most uncertain ones as open class data. We first analyze and compare different uncertainty measurements on the discriminability of various predictions.

Entropy measures the smoothness of the class distribution, which is higher for data in $\overline{\mathcal{C}}^t$ and lower for data in $\mathcal{C}$. We argue that *entropy exhibits low discriminability for highly uncertain and extremely sharp predictions.* Fig. 2(a) shows the value of entropy with respect to the probability of three classes. We can observe that when the probability distribution is close to uniform, i.e. very uncertain, the entropy is insensitive to probability changes. For sharp predictions,



(a) Ternary of entropy (b) Ternary of confidence

Fig. 2. Heatmap of entropy (a) and confidence (b) w.r.t. the probability values of three classes. Each edge is the value range $[0, 1]$. The corner area represents class distributions where one label is very likely, while the center area shows nearly uniform distribution.

the entropy change in few classes is non-negligible. When there are a large number of classes, the relative difference of entropy values between sharp predictions is very small compared with the range of entropy values. For example, with m classes, the entropy values range is $[0, \log(m)]$ but the entropy difference between prediction $(1, 0, 0, ...)$ and $(0.5, 0.5, 0, ...)$ is $\log(2)$. When m is large, such difference can be ignored, but actually, the two predictions are quite different in terms of uncertainty. So estimating the uncertainty only by the entropy will fail to discriminate uncertain and extremely sharp predictions.

Confidence is higher for a more certain data point in $\mathcal{C}$. As shown in Fig. 2(b), confidence value shows ternary contour lines, where the confidence, i.e. the largest probabilities of three classes, is the same. We have the following statement on the length of the contour line: *The contour lines for extremely high and low confidence are short.* The proof is shown in the supplementary. On each contour line, even the confidence of different class distributions are the same, the degrees of uncertainty are different. For example, when the confidence is 0.5, the largest probability is 0.5, and the other two probabilities could be $(0.5, 0)$ or $(0.25, 0.25)$. It is obvious that $(0.5, 0.5, 0)$ is more uncertain than $(0.5, 0.25, 0.25)$. Therefore, confidence lacks discriminability in each contour line. The longer the contour line, the more class distributions in the contour line, the severer the problem of confusing various class distributions. Thus, a shorter contour line exhibits higher discriminability for predictions, which, opposite and complementary to entropy, corresponds to extremely uncertain and confident predictions.

Based on the above analyses, confidence and entropy are complementary to cover both smooth and non-smooth class distributions. However, confidence suffers from prediction errors. If the classifier predicts an open class data as a class in $\mathcal{C}$ with high confidence, the confidence will mistakenly select the data as a common class sample. To compensate confidence, we employ **Consistency** built on multiple diverse classifiers $G_i|_{i=1}^m$, which reflects the agreement of different classifiers. The loss $\mathcal{E}(G_i)$ for the classifier G_i is defined as

$$\mathcal{E}(G_i) = \mathsf{E}_{(\mathbf{x},\mathbf{y})\sim p} L\left(\mathbf{y}, G_i(F(\mathbf{x}))\right) \tag{1}$$

Table 2. Comparison of Calibration Methods on Out-of-distribution Data

Method	Extra Requirement	Extra Computation
Temp Scaling [9]	Target validation set	Training calibration parameters
Dropout [7]	Multiple dropout layers	Multiple full passes
Ensembles [16]	Multiple one-layer classifiers	Multiple one-layer passes
SVI [38]	Several times of model parameters	Several times of computation

where $i = 1, ..., m$ and L is the standard cross-entropy loss. To keep the diversity of different classifiers, we do not back-propagate gradients from $G_i|_{i=1}^m$ to the feature extractor F and initialize G_i with different random initialization. The lower the consistency value, the more likely the data is in $\mathcal{C}$. Consistency is more robust to prediction errors since the probability that all classifiers make the same mistake is low, which means all diverse classifiers predict a sample wrongly and coincidentally into the same class. Therefore, consistency compensates confidence for prediction errors. Confidence usually fails on smooth distribution because they are close to each other and show high consistency though they are uncertain.

Based on the above comparison, we can conclude that entropy, confidence and consistency all have their advantages and drawbacks and cannot individually represent the uncertainty. But they are complementary to each other and can collaborate to form an uncertainty measurement with high discriminability for all types of class distributions. Therefore, we choose the mixture of the three criteria. With each classifier G_i, $(i = 1, ..., m)$ predicting a probability $\hat{\mathbf{y}}_i^*$ for $\mathbf{x}^*, (* = s, t)$ over the source classes $\mathcal{C}^s$, we compute entropy w_{ent}, confidence w_{conf} and consistency w_{cons} as follows:

$$w_{\text{ent}}(\hat{\mathbf{y}}_i^t|_{i=1}^m) = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{j=1}^{|\mathcal{C}^s|} -\hat{y}_{ij}^t \log\left(\hat{y}_{ij}^t\right)\right), \tag{2}$$

$$w_{\text{conf}}(\hat{\mathbf{y}}_i^t|_{i=1}^m) = \frac{1}{m}\sum_{i=1}^{m}\max(\hat{\mathbf{y}}_i^t), \tag{3}$$

$$w_{\text{cons}}(\hat{\mathbf{y}}_i^t|_{i=1}^m) = \frac{1}{|\mathcal{C}^s|}\left\|\frac{1}{m}\sum_{i=1}^{m}\left(\hat{\mathbf{y}}_i^t - \frac{1}{m}\sum_{i=1}^{m}\hat{\mathbf{y}}_i^t\right)^2\right\|_1, \tag{4}$$

where $\hat{y}_{ij}^t$ is the probability of j-th class and max take the maximum entry in $\hat{\mathbf{y}}_i^t$. w_{cons} is the standard deviation of all predictions. Multiple classifiers are employed to calibrate the entropy and the confidence.

We normalize the w_{ent} and w_{cons} by minmax normalization to unify them within $[0, 1]$. Then we compute w^t by aggregating the three uncertainties,

$$w^t = \frac{(1 - w_{\text{ent}}) + (1 - w_{\text{cons}}) + w_{\text{conf}}}{3}, \tag{5}$$

where the higher the $w_t(\mathbf{x}_0^t)$, the more likely $\mathbf{x}_0^t$ is in $\mathcal{C}$.

For source weight, since our novel w^t can distinguish target private data from common data more clearly and common class data should have high probability on one of the common classes, we sum the prediction of common data that are selected by w^t to compute weights $\mathbf{V}$ for source classes. Such class-level weight is only high for source classes in $\mathcal{C}$. Since source labels are available, the source weight w^s can be easily defined by taking the $\mathbf{y}^s$-th class weight:

$$\mathbf{V} = \text{avg}_{w^t(x^t)>w_0}\, \hat{\mathbf{y}}^t \text{ and } w^s(x^s) = V_{\mathbf{y}^s}, \tag{6}$$

where avg computes the average of $\hat{\mathbf{y}}^t$ and $\mathbf{V}_{\mathbf{y}^s}$ is the $\mathbf{y}^s$-th entry of $\mathbf{V}$.

3.3 Uncertainty Calibration

The transferability introduced above can estimate the uncertainty for all types of predictions to detect open classes. However, the criterion is still not reliable enough for UniDA. As shown in [16], overconfident predictions with low uncertainties exist among data of the "unknown" class, i.e., out-of-distribution data. So the uncertainty estimated from the prediction does not reflect the real uncertainty of the data samples, which deteriorates the reliability of the transferability.

Calibration is a widely-used approach to estimate the uncertainty more accurately, so we employ the most suitable calibration method for deep UniDA, where large-scale parameters and the domain gap need consideration. We compare existing calibration methods surveyed in [32]: Vanilla, Temp Scaling [9], Dropout [7], Ensemble [16], SVI [38], in terms of performance on out-of-distribution data. We do not include LL for the low performance in [32] and extra Bayesian Network.

As shown in Table.2, Temp Scaling requires a target validation set, which is not available in UniDA, or otherwise we know the components of the "unknown" class. Dropout and SVI require far more computation on deep networks. SVI can be embedded into particular network architectures. Ensemble naturally utilizes the current multi-classifier architecture in our framework and introduce no extra computation. From [32], we observe that when *testing on out-of-distribution data, Ensemble achieves the best performance on large-scale datasets*. Thus, Ensemble is the most suitable framework for UniDA and already embedded in our framework.

3.4 Calibrated Multiple Uncertainties Framework

We first introduce the framework of CMU, which is shown in Fig. 3. CMU consists of a feature extractor F, a label classifier G, multiple classifiers G_1, G_2 ... G_m, and a domain discriminator D. For a data point $\mathbf{x}$, F extracts the feature $\mathbf{z} = F(\mathbf{x})$ and G predicts a probability $\hat{\mathbf{y}} = G(\mathbf{z})$ for $\mathbf{x}$. We derive the transferability measurement w^s and w^t for source and target data from the output of multiple classifiers as Equation (5) and (6), which is used to weight each data sample in distribution matching. We train the F, G on D as [8] to enable classification and feature distribution matching where the losses are defined as:

$$\mathcal{E}(G) = \mathsf{E}_{(\mathbf{x},\mathbf{y})\sim p} L\left(\mathbf{y}, G(F(\mathbf{x}))\right) \tag{7}$$

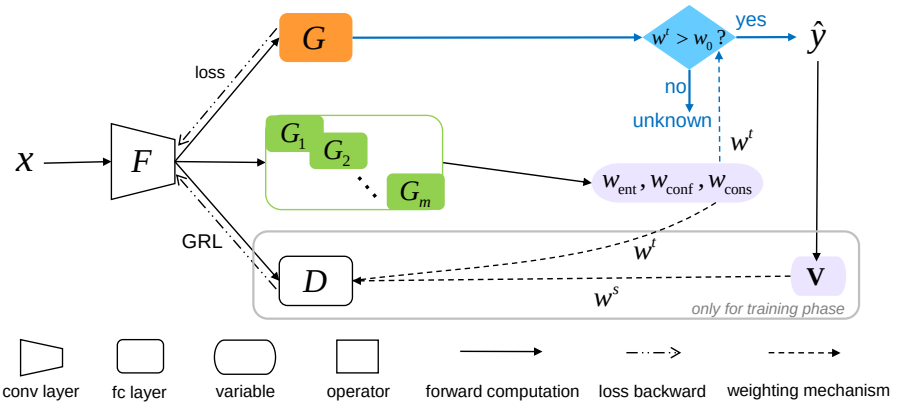


Fig. 3. The architecture of Calibrated Multiple Uncertainties (CMU). An input $\mathbf{x}$ is fed to F to output a feature, which is then input to a classifier G for prediction. The feature is also input to m classifiers $G_i|_{i=1}^m$ for ensemble. Three uncertainties: entropy, consistency and confidence are defined on the output of $G_i|_{i=1}^m$, to produce target weight w^t, which is used to decide a target data is "unknown" or not. The blue solid/dot lines represent the mechanism judging the "unknown" class. Target data with $w^t \geq w_0$ is common class data and is given a class prediction, and otherwise it is classified as "unknown". The source weight w^s is derived from class-level weight $\mathbf{V}$ based on the prediction $\hat{y}$ of target common class data. w^s and w^t are used to weighting samples in distribution matching. The part in the gray square is only used in the training phase.

$$\mathcal{E}(D) = -\,\mathsf{E}_{\mathbf{x}\sim p} w^s(\mathbf{x}) \log D\left(F\left(\mathbf{x}\right)\right) - \mathsf{E}_{\mathbf{x}\sim q} w^t(\mathbf{x}) \log\left(1 - D\left(F\left(\mathbf{x}\right)\right)\right) \tag{8}$$

where L is the cross-entropy loss. Combined with the loss for multiple classifiers in (1), the optimization of the new architecture can be defined as follows,

$$\begin{aligned} &\max_{D} \min_{F,G} \mathcal{E}(G) - \lambda \mathcal{E}(D) \\ &\min_{G_i|_{i=1}^m} \sum_{i=1}^{m} \mathcal{E}(G_i). \end{aligned} \tag{9}$$

In the testing phase, given each input target sample $\mathbf{x}_0$, we first compute $w^t(\mathbf{x}_0)$ and then predict the class of $y(\mathbf{x})$ with a validated threshold w_0 as:

$$y(\mathbf{x}_0) = \begin{cases} \texttt{unknown} & w^t(\mathbf{x}_0) \leq w_0 \\ \operatorname{argmax}\left(\hat{\mathbf{y}}_0\right) & w^t(\mathbf{x}_0) > w_0 \end{cases} \tag{10}$$

which either rejects the $\mathbf{x}_0$ as "unknown" class or classifies it to a common class.

Our new transferability measurement consists of three complementary uncertainties covering all class distributions. We carefully compare and employ the most suitable calibration method to improve the quality of the uncertainty estimation. The proposed calibrated multiple uncertainties (CMU) can discriminate target data in $\bar{\mathcal{C}}^t$ from target data in $\mathcal{C}$ more clearly, which in turn helps discriminate source data in $\bar{\mathcal{C}}^s$ from source data in $\mathcal{C}$. Thus, CMU can simultaneously match

the distributions of common classes and detect samples from open classes, which achieves high performance on both classifying common class and open class data.

4 Experiments

We conduct a thorough evaluation of CMU on universal domain adaptation benchmarks. Code is at https://github.com/thuml/Calibrated-Multiple-Uncertainties.

4.1 Setup

Datasets We perform experiments on **Office-31** [28], **Office-Home** [35], **VisDA** [27] and **DomainNet** [26] datasets. For the first three datasets, we follow the same setup as [41]. **DomainNet** is by far the largest domain adaptation dataset, consists of six distinct domains: Clipart(C), Infograph(I), Painting(P), Quickdraw(Q), Real(R) and Sketch(S) across 345 classes. In the alphabet order, we use the first 150 classes as $\mathcal{C}$, the next 50 classes as $\overline{\mathcal{C}}^s$ and the rest as $\overline{\mathcal{C}}^t$. We choose 3 domains to transfer between each other due to the large amount of data.

Compared Methods. We compare the proposed CMU with (1) ResNet [10], (2) close-set domain adaptation: Domain-Adversarial Neural Networks (DANN) [8], Residual Transfer Networks (RTN)[22], (3) partial domain adaptation: Importance Weighted Adversarial Nets (IWAN) [44], Partial Adversarial Domain Adaptation (PADA) [3], (4) open set domain adaptation: Assign-and-Transform-Iteratively (ATI) [25], Open Set Back-Propagation (OSBP) [30]. (5)universal domain adaptation: Universal Adaptation Network (UAN) [41].

Evaluation Protocols Previous work [41] uses the per-class accuracy as the evaluation metric, which calculates the instance accuracy of each class and then average. However, in per-class accuracy, the accuracy of each common class has the same contribution as the whole "unknown" class. So the influence of the "unknown" class is small, especially when the amount of common classes is large. As shown in Fig. 1(b), with a large number of classes, only classifying common class samples correctly can achieve fairly high per-class accuracy. Inspired by the F1-score, we propose the **H-score**: the harmonic mean of the instance accuracy on common class $a_{\mathcal{C}}$ and accuracy on the "unknown" class $a_{\overline{\mathcal{C}}^t}$ as:

$$h = 2 \cdot \frac{a_{\mathcal{C}} \cdot a_{\overline{\mathcal{C}}^t}}{a_{\mathcal{C}} + a_{\overline{\mathcal{C}}^t}}. \tag{11}$$

The new evaluation metric is high only when both the $a_{\mathcal{C}}$ and $a_{\overline{\mathcal{C}}^t}$ are high. So H-score emphasizes the importance of both abilities of UniDA methods.

Implementation Details We implement our method in PyTorch framework with ResNet-50 [10] backbone pretrained on ImageNet [6]. The hyperparameters are tuned with cross-validation [42] and fixed for each dataset. To enable more diverse classifiers in deep ensemble, we use different data augmentations and randomly shuffled data in different orders for different classifiers.

Table 3. Average class accuracy (%) and H-score (%) on **Office-31**

Method	Office-31													
	A → W		D → W		W → D		A → D		D → A		W → A		Avg	
	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score
ResNet [10]	75.94	47.92	89.60	54.94	90.91	55.60	80.45	49.78	78.83	48.48	81.42	48.96	82.86	50.94
DANN [8]	80.65	48.82	80.94	52.73	88.07	54.87	82.67	50.18	74.82	47.69	83.54	49.33	81.78	50.60
RTN [22]	85.70	50.21	87.80	54.68	88.91	55.24	82.69	50.18	74.64	47.65	83.26	49.28	83.83	51.21
IWAN [44]	85.25	50.13	90.09	54.06	90.00	55.44	84.27	50.64	84.22	49.65	86.25	49.79	86.68	51.62
PADA [44]	85.37	49.65	79.26	52.62	90.91	55.60	81.68	50.00	55.32	42.87	82.61	49.17	79.19	49.98
ATI [25]	79.38	48.58	92.60	55.01	90.08	55.45	84.40	50.48	78.85	48.48	81.57	48.98	84.48	51.16
OSBP [30]	66.13	50.23	73.57	55.53	85.62	57.20	72.92	51.14	47.35	49.75	60.48	50.16	67.68	52.34
UAN [41]	85.62	58.61	94.77	70.62	97.99	71.42	86.50	59.68	85.45	60.11	85.12	60.34	89.24	63.46
CMU	**86.86**	**67.33**	**95.72**	**79.32**	**98.01**	**80.42**	**89.11**	**68.11**	**88.35**	**71.42**	**88.61**	**72.23**	**91.11**	**73.14**

Table 4. Tasks on **DomainNet** and **VisDA** dataset

Method	DomainNet (H-score)							VisDA	
	P → R	R → P	P → S	S → P	R → S	S → R	Avg	Acc	H-score
ResNet [10]	30.06	28.34	26.95	26.95	26.89	29.74	28.15	52.80	25.44
DANN [8]	31.18	29.33	27.84	27.84	27.77	30.84	29.13	52.94	25.65
RTN [22]	32.27	30.29	28.71	28.71	28.63	31.90	30.08	53.92	26.02
IWAN [44]	35.38	33.02	31.15	31.15	31.06	34.94	32.78	58.72	27.64
PADA [44]	28.92	27.32	26.03	26.03	25.97	28.62	27.15	44.98	23.05
ATI [25]	32.59	30.57	28.96	28.96	28.89	32.21	30.36	54.81	26.34
OSBP [30]	33.60	33.03	30.55	30.53	30.61	33.65	32.00	30.26	27.31
UAN [41]	41.85	43.59	39.06	38.95	38.73	43.69	40.98	60.83	30.47
CMU	**50.78**	**52.16**	**45.12**	**44.82**	**45.64**	**50.97**	**48.25**	**61.42**	**34.64**

4.2 Results

The classification results of Office-31, VisDA, Office-Home and DomainNet are shown in Table 3, 4 and 5. For a fair comparison with UAN, we compute per-class accuracy on Office-31 and VisDA. CMU outperforms UAN and all other methods. We compare H-score on all datasets and CMU consistently outperforms previous methods with a large margin on all datasets with various difficulties of detecting open classes. Some domain adaptation methods for other settings perform even worse than ResNet due to the violation of the label space assumption.

In particular, UAN performs well on per-class accuracy but not well on H-score, because the sub-optimal transferability measurement of UAN causes it unable to detect open classes clearly. The low accuracy of the "unknown" class pulls down the H-score. CMU outperforms UAN on H-score with a large margin, which demonstrates that CMU has higher-quality transferability measurement to more accurately detect target open classes $\overline{\mathcal{C}}^t$. This boosts the accuracy of the "unknown" class and further improves the quality of w^s, which further constrains feature distribution alignment within $\mathcal{C}$ and improves the common class accuracy.

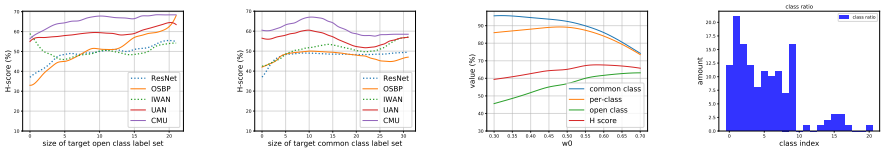
Table 5. H-score (%) of tasks on on **Office-Home** dataset

Method	Office-Home												
	A→C	A→P	A→R	C →A	C→P	C→R	P→A	P→C	P→R	R→A	R→ C	R→P	Avg
ResNet [10]	44.65	48.04	50.13	46.64	46.91	48.96	47.47	43.17	50.23	48.45	44.76	48.43	47.32
DANN [8]	42.36	48.02	48.87	45.48	46.47	48.37	45.75	42.55	48.70	47.61	42.67	47.40	46.19
RTN [22]	38.41	44.65	45.70	42.64	44.06	45.48	42.56	36.79	45.50	44.56	39.79	44.53	42.89
IWAN [44]	40.54	46.96	47.78	44.97	45.06	47.59	45.81	41.43	47.55	46.29	42.49	46.54	45.25
PADA [44]	34.13	41.89	44.08	40.56	41.52	43.96	37.04	32.64	44.17	43.06	35.84	43.35	40.19
ATI [25]	39.88	45.77	46.63	44.13	44.39	46.63	44.73	41.20	46.59	45.05	41.78	45.45	44.35
OSBP [30]	39.59	45.09	46.17	45.70	45.24	46.75	45.26	40.54	45.75	45.08	41.64	46.90	44.48
UAN [41]	51.64	51.7	54.3	61.74	57.63	61.86	50.38	47.62	61.46	62.87	52.61	65.19	56.58
CMU	**56.02**	**56.93**	**59.15**	**66.95**	**64.27**	**67.82**	**54.72**	**51.09**	**66.39**	**68.24**	**57.89**	**69.73**	**61.60**

4.3 Analysis

Varying Size of $\overline{\mathcal{C}}^s$ and $\overline{\mathcal{C}}^t$ Following UAN, with fixed $|\mathcal{C}^s \cup \mathcal{C}^t|$ and $|\mathcal{C}^s \cap \mathcal{C}^t|$, we explore the H-score with the various sizes of $\overline{\mathcal{C}}^t$ ($\overline{\mathcal{C}}^s$ also changes correspondingly) on task A → D in Office-31 dataset. As shown in Figure 4(a), CMU outperforms all the compared methods consistently with different $\overline{\mathcal{C}}^t$, proving that CMU is effective and robust to diverse $\overline{\mathcal{C}}^s$ and $\overline{\mathcal{C}}^t$. In particular, when $\overline{\mathcal{C}}^t$ is large (over 10), meaning there are many open classes, CMU outperforms other methods with a large margin, demonstrating that CMU is superior in detecting open classes.

Varying Size of Common Label $\mathcal{C}$ Following UAN, we fix $|\mathcal{C}^s \cup \mathcal{C}^t|$ and varying $\mathcal{C}$ on task A → D in Office-31 dataset. We let $|\overline{\mathcal{C}}^s| + 1 = |\overline{\mathcal{C}}^t|$ to keep the relative size of $\overline{\mathcal{C}}^s$ and $\overline{\mathcal{C}}^t$ and vary $\mathcal{C}$ from 0 to 31. As shown in Figure 4(b), CMU consistently outperforms previous methods on all size of $\mathcal{C}$. In particular, when the source domain and the target domain have no overlap on label sets, all the target data should be marked as "unknown". CMU achieves much higher H-score, indicating that CMU can detect open classes more effectively. When $|\mathcal{C}| = 31$, the setting degrades to closed set domain adaptation, CMU and UAN perform similarly, because there is no open class to influence the adaptation.



(a) H-score w.r.t. $\overline{\mathcal{C}}^t$ (b) H-score w.r.t. $\mathcal{C}$ (c) Metric compare (d) Class ratio

Fig. 4. (a)(b) H-score with respect to $\overline{\mathcal{C}}^t$ and $\mathcal{C}$. In (a), we fix $|\mathcal{C}^s \cup \mathcal{C}^t|$ and $|\mathcal{C}^s \cap \mathcal{C}^t|$; In (b), we fix $|\mathcal{C}^s \cup \mathcal{C}^t|$. (c) Relationship between different metrics and w_0. (d) The class ratio of predicted labels (used to compute w^s) of target data in all source labels. Classes 0-9 are source commons and 10-19 are source privates.

Ablation Study We go deeper into the efficacy of the proposed method by evaluating variants of CMU on Office-31. (1) CMU w/o cons is the variant without using the consistency component in the uncertainty in Eq. (4) but still using multiple classifiers to calibrate the entropy and confidence; (2) CMU w/o conf is the variant without integrating the average confidence of classifier in Eq. (3). (3) CMU w/o ent is the variant without integrating the average entropy of classifier into the criterion in Eq. (2). (4) CMU w/o ensemble is the variant without calibrating entropy and confidence but still using single classifier G to compute entropy and confidence while multiple classifiers are still used to compute consistency. (5) CMU w/ domain sim is the variant by adding the domain similarity as another component in the transferability like UAN [41].

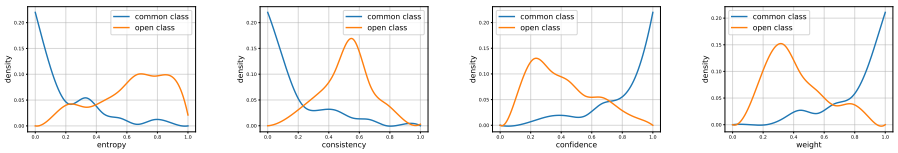
Table 6. Ablation Study tasks on **Office-31** dataset

Method	D → W		A → D		W → A		Avg (6 task)	
	Acc	H-score	Acc	H-score	Acc	H-score	Acc	H-score
CMU	95.72	79.32	89.11	68.11	88.61	72.23	91.11	73.14
w/o cons	95.01	78.65	88.74	67.25	87.82	71.44	90.43	72.23
w/o conf	95.23	78.84	88.92	67.48	88.04	71.71	90.62	72.52
w/o ent	94.11	75.68	86.81	63.97	87.24	68.66	89.07	69.78
w/o ensemble	93.68	74.43	86.39	63.81	88.67	72.26	88.93	69.50
w/ domain sim	95.70	79.30	89.63	68.14	88.67	72.26	91.28	73.15

As shown in Table 6, CMU outperforms CMU w/o cons/conf/ent, especially w/o entropy, indicating the contribution of the multiple uncertainties is complementary to achieve a more complete and accurate uncertainty estimation. CMU outperforms CMU w/o ensemble, proving the calibration from the ensemble can more accurately estimate the uncertainty. CMU w/ domain sim performs similarly to CMU, indicating that domain similarity has little effect on detecting open classes, and thus we do not include it in the uncertainty estimation.

Comparison of Multiple Metrics To justify our new H-score, we visualize the relationship between different metrics w.r.t. w_0 in Figure 4(c). We can observe that the open class accuracy increases with w_0 increasing while the common class accuracy decreases with w_0 increasing. This is because, with higher w_0, more data are marked as "unknown" and more common data are misclassified to "unknown". Per-class accuracy varies in the same trend of common class accuracy, indicating that per-class accuracy bias common class accuracy while nearly neglect the open class accuracy. H-score is high only when both common class and "unknown" accuracies are high, which more comprehensively evaluates UniDA methods.

Threshold sensitivity We investigate the sensitivity of CMU with respect to threshold w_0 in task A → D. As shown in Figure 4(c), with w_0 varying in a reasonable range $[0.45, 0.60]$, the H-score changes little, which proves that the performance is not very sensitive to the threshold w_0.



(a) entropy (b) Consistency (c) confidence (d) final weight

Fig. 5. Density of each criterion within common and open class data.

Hypotheses Justification To justify that our new transferability measurement distinguishes data in the common label set from those in the private label sets, we plot the estimated probability density function for different components of weights $w^s(x)$ in Eq. (6) and $w^t(x)$ in Eq. (5) on A $\rightarrow$ D task of Office-31. Figure 5(a)-5(d) show that the three uncertainties: entropy, consistency and confidence all distinguish target data in $\mathcal{C}$ and $\overline{\mathcal{C}}^t$ clearly, proving that the multi-classifier ensemble model can calibrate the uncertainty and estimate it more accurately. Figure 4(d) (0-9 is the common class) proves that the source class-level weight could assign high weights for common classes, which in turn demonstrates that the selected data to compute source weight are mostly common classes.

5 Conclusion

In this paper, we propose a novel approach: Calibrated Multiple Uncertainties (CMU) and a new evaluation metric: H-score for Universal Domain Adaptation (UniDA). We design a novel transferability consisting of entropy, confidence and consistency, calibrated by a deep ensemble model. The new transferability exploits complementary characteristics of different uncertainties to cover all types of predictions. The calibration more accurately estimates the uncertainty and improves the quality of the transferability. The advanced transferability, in turn, improves the quality of source weight. CMU achieves a balanced ability to detect open classes and classify common class data correctly. We further propose a novel H-score to compensate for the previous per-class accuracy for ignorance of open classes. A thorough evaluation shows that CMU outperforms the state-of-the-art UniDA method on both the common set accuracy and the "unknown" class accuracy, especially with a large margin on detecting open classes.

Acknowledgement

This work was supported by the Natural Science Foundation of China (61772299, 71690231), and China University S&T Innovation Plan Guided by the Ministry of Education.

References

1. Busto, P.P., Iqbal, A., Gall, J.: Open set domain adaptation for image and action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
2. Cao, Z., Long, M., Wang, J., Jordan, M.I.: Partial transfer learning with selective adversarial networks. In: CVPR (June 2018)
3. Cao, Z., Ma, L., Long, M., Wang, J.: Partial adversarial domain adaptation. In: ECCV. pp. 135–150 (2018)
4. Cao, Z., You, K., Long, M., Wang, J., Yang, Q.: Learning to transfer examples for partial domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2985–2994 (2019)
5. Chen, Q., Liu, Y., Wang, Z., Wassell, I., Chetty, K.: Re-weighted adversarial adaptation network for unsupervised domain adaptation. In: CVPR. pp. 7976–7985 (2018)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
7. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059 (2016)
8. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.S.: Domain-adversarial training of neural networks. JMLR **17**, 59:1–59:35 (2016)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1321–1330. JMLR. org (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. Hu, J., Wang, C., Qiao, L., Zhong, H., Jing, Z.: Multi-weight partial domain adaptation. In: The British Machine Vision Conference (BMVC) (2019)
12. Hu, L., Kan, M., Shan, S., Chen, X.: Duplex generative adversarial network for unsupervised domain adaptation. In: CVPR (June 2018)
13. Hu, Y., Stumpfe, D., Bajorath, J.: Computational exploration of molecular scaffolds in medicinal chemistry: Miniperspective. Journal of medicinal chemistry **59**(9), 4062–4076 (2016)
14. Kang, G., Zheng, L., Yan, Y., Yang, Y.: Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. In: ECCV (September 2018)
15. Konstantinos, B., Nathan, S., David, D., Dumitru, E., Dilip, K.: Unsupervised pixel–level domain adaptation with generative adversarial networks. In: CVPR. pp. 95–104 (2017)
16. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems. pp. 6402–6413 (2017)
17. Lian, Q., Li, W., Chen, L., Duan, L.: Known-class aware self-ensemble for open set domain adaptation. arXiv preprint arXiv:1905.01068 (2019)
18. Liu, H., Cao, Z., Long, M., Wang, J., Yang, Q.: Separate to adapt: Open set domain adaptation via progressive separation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
19. Liu, Y.C., Yeh, Y.Y., Fu, T.C., Wang, S.D., Chiu, W.C., Frank Wang, Y.C.: Detach and adapt: Learning cross-domain disentangled deep representation. In: CVPR (June 2018)

20. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: ICML (2015)
21. Long, M., Cao, Z., Wang, J., Jordan, M.I.: Conditional domain adversarial network. In: NeurIPS (2018)
22. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. In: NeurIPS. pp. 136–144 (2016)
23. Maria Carlucci, F., Porzi, L., Caputo, B., Ricci, E., Rota Bulo, S.: Autodial: Automatic domain alignment layers. In: ICCV (Oct 2017)
24. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: CVPR (June 2018)
25. Panareda Busto, P., Gall, J.: Open set domain adaptation. In: ICCV (Oct 2017)
26. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1406–1415 (2019)
27. Peng, X., Usman, B., Kaushik, N., Wang, D., Hoffman, J., Saenko, K., Roynard, X., Deschaud, J.E., Goulette, F., Hayes, T.L.: VisDA: A synthetic-to-real benchmark for visual domain adaptation. In: CVPR Workshops. pp. 2021–2026 (2018)
28. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: ECCV (2010)
29. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: CVPR (June 2018)
30. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: ECCV (September 2018)
31. Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R.: Generate to adapt: Aligning domains using generative adversarial networks. In: CVPR (June 2018)
32. Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J., Ren, J., Nado, Z.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In: Advances in Neural Information Processing Systems. pp. 13969–13980 (2019)
33. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR (2017)
34. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
35. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: CVPR (2017)
36. Vittorio, F., Alina, K., Rodrigo, B., Victor, G., Matteo, M.: Open images challenge 2019. `https://storage.googleapis.com/openimages/web/challenge2019.html` (2019)
37. Volpi, R., Morerio, P., Savarese, S., Murino, V.: Adversarial feature augmentation for unsupervised domain adaptation. In: CVPR (June 2018)
38. Wu, A., Nowozin, S., Meeds, E., Turner, R., Hernández-Lobato, J., Gaunt, A.: Deterministic variational inference for robust bayesian neural networks. In: 7th International Conference on Learning Representations, ICLR 2019 (2019)
39. Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K., Pande, V.: Moleculenet: a benchmark for molecular machine learning. Chemical science **9**(2), 513–530 (2018)
40. Xie, S., Zheng, Z., Chen, L., Chen, C.: Learning semantic representations for unsupervised domain adaptation. In: ICML. pp. 5423–5432 (2018)
41. You, K., Long, M., Cao, Z., Wang, J., Jordan, M.I.: Universal domain adaptation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

42. You, K., Wang, X., Long, M., Jordan, M.: Towards accurate model selection in deep unsupervised domain adaptation. In: ICML. pp. 7124–7133 (2019)
43. Zhang, H., Li, A., Han, X., Chen, Z., Zhang, Y., Guo, Y.: Improving open set domain adaptation using image-to-image translation. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). pp. 1258–1263. IEEE (2019)
44. Zhang, J., Ding, Z., Li, W., Ogunbona, P.: Importance weighted adversarial nets for partial domain adaptation. In: CVPR (June 2018)
45. Zhang, W., Ouyang, W., Li, W., Xu, D.: Collaborative and adversarial network for unsupervised domain adaptation. In: CVPR (June 2018)