Confidence Score for Source-Free Unsupervised Domain Adaptation

Jonghyun Lee [1] **Dahuin Jung** [1] **Junho Yim** [2] **Sungroh Yoon** [1 3]

Abstract

Source-free unsupervised domain adaptation (SFUDA) aims to obtain high performance in the unlabeled target domain using the pre-trained source model, not the source data. Existing SFUDA methods assign the same importance to all target samples, which is vulnerable to incorrect pseudo-labels. To differentiate between sample importance, in this study, we propose a novel sample-wise confidence score, the Joint Model-Data Structure (JMDS) score for SFUDA. Unlike existing confidence scores that use only one of the source or target domain knowledge, the JMDS score uses both knowledge. We then propose a Confidence score Weighting Adaptation using the JMDS (CoWA-JMDS) framework for SFUDA. CoWA-JMDS consists of the JMDS scores as sample weights and weight Mixup that is our proposed variant of Mixup. Weight Mixup promotes the model make more use of the target domain knowledge. The experimental results show that the JMDS score outperforms the existing confidence scores. Moreover, CoWA-JMDS achieves state-of-the-art performance on various SFUDA scenarios: closed, open, and partial-set scenarios.

1. Introduction

Recently, Deep Neural Networks (DNNs) (LeCun et al., 2015) have successfully demonstrated high performance in various applications. However, if the distribution of the training and test data differs, significant performance degradation occurs, which is known as a domain shift (Pan & Yang, 2009). Unsupervised domain adaptation (UDA) mitigates the domain shift problem using both fully annotated source and unlabeled target data with the assumption that the data distributions in the two domains are slightly different. The UDA task aims to obtain a high target performance using the two domains without the target label.

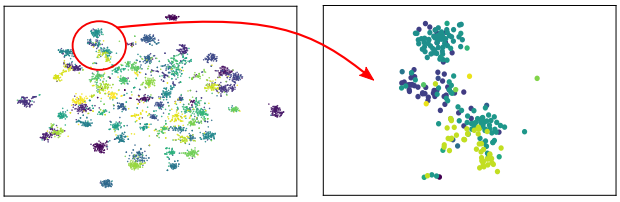


Figure 1. (Left) t-SNE plot of the target feature in Ar → Rw task on the Office-Home dataset. (Right) Zoomed-in t-SNE plot. The color of samples indicates the ground truth labels.

All conventional UDA methods assume the availability of both the source data and corresponding labels. However, this may be impractical in some cases. First, growing concerns regarding data privacy and security force companies to only release the data. Second, many resources such as GPUs and time are required to train a model when source data are much greater than target data. To address these concerns, source-free UDA (SFUDA) has recently been studied (Liang et al., 2020a; Li et al., 2020; Yang et al., 2020; Hou & Zheng, 2021). Instead of accessing the source data, SFUDA assumes that we can access only a pre-trained model using labeled source domain data.

Existing SFUDA methods strictly follow the cluster assumption (Grandvalet et al., 2005) and use pseudo-labels (Lee et al., 2013) of target data based on the target feature cluster. Because the only data accessible in SFUDA are the target data, the model is updated while preserving its intrinsic cluster architecture. In other words, existing SFUDA methods train the model so that its decision boundary does not penetrate the target feature cluster. Figure 1 shows t-SNE (Van der Maaten & Hinton, 2008) plots to visualize the inherent architecture of the target features obtained by the pre-trained source model. As shown in Figure 1, the samples form their own cluster. However, according to the right side of Figure 1, ground truth labels of samples included in the same cluster are diverse for some clusters. The model assigns the same pseudo-label to samples in the same cluster based on the cluster assumption, which leads incorrect pseudo-labels. To robustly learn with incorrect pseudo-labels, samples with low-confidence in their pseudo-label should be suppressed when the model trains (Ding et al., 2020). The existing methods are limited in that they

[1]Data Science and AI Lab., Seoul National University [2]AIRS Company, Hyundai Motor Group, Seoul, Korea [3] Department of ECE and Interdisciplinary Program in AI, Seoul National University. Correspondence to: Sungroh Yoon <sryoon@snu.ac.kr>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

are vulnerable to incorrect pseudo-labels and confirmation bias (Arazo et al., 2020), also known as noise accumulation, because it allocates the same weight to all samples regardless of their confidence in the pseudo-labels.

In this study, we propose a novel confidence score for SFUDA, the Joint Model-Data Structure (JMDS) score, to differentiate between sample importance based on the confidence for pseudo-labels. SFUDA has two components: the pre-trained source model and target data. The model has knowledge of the source domain, such as class similarity, because it was trained in the source domain. On the other hand, the target feature distribution obtained from the target data has knowledge of the target domain such as data similarity. Therefore, the JMDS score should include knowledge of the model and target feature distribution to fully utilize knowledge of both domains.

The JMDS score consists of two confidence scores, a Log Probability Gap (LPG) score and a Model Probability of Pseudo-Label (MPPL) score, to include both knowledge. We first use Gaussian Mixture Modeling (GMM) in the target feature space to obtain the log-likelihood and pseudo-label of each sample. The LPG score, the data-structure-wise confidence score, is the gap between the primary and secondary classes of log probability based on GMM. The MPPL score, the model-wise confidence score, is the probability of the model for a corresponding pseudo-label obtained from GMM. The LPG and MPPL scores include knowledge of the target and source domains, respectively. Therefore, to the best of our knowledge, the JMDS score is the first confidence score that includes both source and target domain knowledge.

The objective of the JMDS score is to measure the sample-wise confidence for pseudo-labels at the given model, not the adaptation. To use the JMDS score in the learning process for SFUDA, we propose a novel framework, Confidence Score Weighting Adaptation using the JMDS (CoWA-JMDS) framework. CoWA-JMDS uses the JMDS score as a sample-wise weight and pseudo-labels obtained from GMM. Also, it uses weight Mixup, which is our proposed variant of Mixup (Zhang et al., 2017). Because sample weighting with the JMDS score suppresses low-confidence samples, knowledge of the target feature distribution may not be sufficiently exploited. Weight Mixup promotes the model make more use of target domain knowledge by mixing low-confidence samples with other samples and considering confidence of the mixed samples. CoWA-JMDS can be easily extended to open-set and partial-set scenarios with minor modifications.

We evaluated the JMDS score and CoWA-JMDS on various public UDA benchmarks. First, the JMDS score achieved the best performance in terms of measuring confidence compared to existing confidence scores. Because performance using the JMDS is better than using the MPPL or LPG alone, it has been experimentally proven that knowledge of both domains is important in SFUDA. Second, despite its simplicity, CoWA-JMDS outperformed the state-of-the-art SFUDA method on three benchmarks in closed-set scenarios without any auxiliary networks. CoWA-JMDS also achieved the best performance in open-set and partial-set scenarios. Through further analysis in Section 6, we provide an explanation on why the JMDS score is reliable in SFUDA. Additionally, our proposed weight Mixup improves 3.4% in terms of the average accuracy on Office-31 dataset than Mixup. It demonstrates that weight Mixup is an effective technique in learning with sample weighting based on confidence scores. The code is available at https://github.com/Jhyun17/CoWA-JMDS. Our contributions can be summarized as follows:

- We propose a novel confidence score, the JMDS score, which considers knowledge of both the source and target domains in SFUDA.
- We propose an SFUDA framework, CoWA-JMDS, which uses JMDS scores as sample-wise weights and weight Mixup which is proposed to exploit more target domain knowledge.
- We demonstrate that the proposed JMDS score and CoWA-JMDS achieve state-of-the-art performance on UDA benchmarks.

2. Related work

2.1. Source-free unsupervised domain adaptation

SFUDA is a more difficult UDA setting than the conventional UDA, where only a source-trained model can be used. Thus, the existing methods for UDA that use source data directly cannot be applied to SFUDA.

Collaborative Class Conditional Generative Adversarial Networks (3C-GAN) (Li et al., 2020) are based on time-consuming target-style image generation through a conditional GAN. Source-Free Image Translation(SFIT) (Hou & Zheng, 2021) uses knowledge distillation to translate target images into source style without using source images. Both methods require an auxiliary network.

Source HypOthesis Transfer (SHOT) (Liang et al., 2020a) matches the target feature to a fixed pre-trained source classifier that fine-tunes the feature extractor. SHOT uses self-supervised pseudo-labeling (SSPL) and information maximization loss to balance the pseudo-label and avoid trivial solutions. However, SHOT cannot fully extract the knowledge of the target feature structure because it uses SSPL which does not consider the covariance of each dimension on the feature space and cluster density.

Neighborhood Reciprocity Clustering (NRC) (Yang et al., 2021) exploits the intrinsic neighborhood structure of the

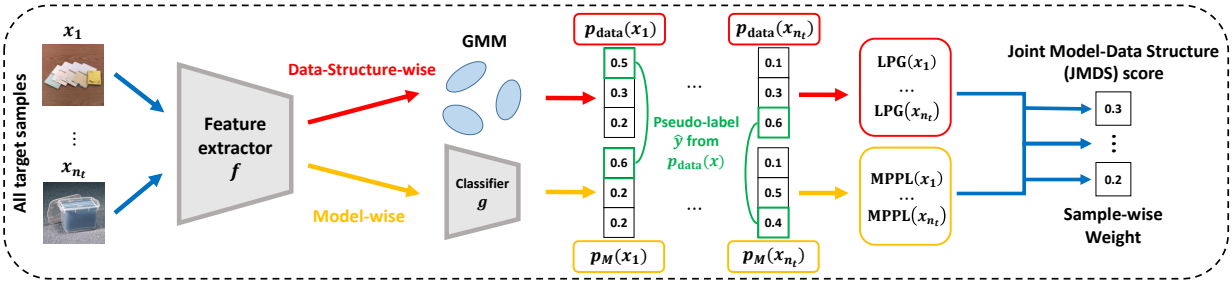


Figure 2. Overview of the JMDS score. The JMDS score consists of two scores: The LPG and MPPL scores. The LPG score uses the data-structure-wise probability from GMM, while the MPPL score uses the model-wise probability. The red color indicates the data-structure-wise knowledge and the yellow color indicates the model-wise knowledge. The pseudo-labels of samples are decided by the data-structure-wise probability $p_{\text{data}}(X_t)$.

target data in the feature space. It uses the nearest neighbors and an affinity matrix on the target feature space to exploit the knowledge of the target data distribution. However, it is difficult to use existing techniques such as Mixup (Zhang et al., 2017) and data augmentation.

In this study, we propose an SFUDA framework to overcome these weaknesses. It uses GMM on the feature space to extract knowledge of the target data structure. GMM has the advantage of obtaining a sample log-likelihood compared to other clustering methods. Additionally, the framework does not require auxiliary networks and can be effectively combined with a variant of Mixup.

2.2. Confidence score

Geifman et al. (2018) divided confidence scores into two main tasks; ordinal ranking and probability calibration. In this study, we focused on ordinal rankings. Ordinal ranking is commonly used for selective classification (Lakshminarayanan et al., 2016; Geifman & El-Yaniv, 2017; Mandelbaum & Weinshall, 2017; Nair et al., 2020), which is a task that discriminates samples according to their confidence level for labels to avoid low-confidence samples during training. In this work, we present for the first time an SFUDA method that uses a confidence score to robustly learn low-confidence samples that are more likely to have incorrect pseudo-labels.

3. Joint Model-Data Structure (JMDS) score

In SFUDA, we can only use the target data $X_t = \{x_i^t\}_{i=1}^{n_t}$ and model M, not the source data. The ground truth labels of the target data, $Y_t = \{y_i^t\}_{i=1}^{n_t}$, are inaccessible during the learning stage. Instead, the pseudo-labels of the target data, $\hat{Y}_t = \{\hat{y}_i^t\}_{i=1}^{n_t}$, are utilized during the learning stage. The pseudo-labels are obtained through the prediction of the model or the feature distribution. The model M is pre-trained using labeled source data $X_s = \{x_i^s.y_i^s\}_{i=1}^{n_s}$. Here, M is composed of a feature extractor $f : X \to \mathbb{R}^d$ and a classifier $g : \mathbb{R}^d \to \mathbb{R}^K$ where d is the dimension of the feature and K is the number of classes. Given the probability $p(x) = \big(p(x)_1, p(x)_2, \cdots, p(x)_K\big)$ where $p(x)_k = p(y = k|x)$, the model-wise probability $p_M(X_t)$ is expressed as follows:

$$p_M(X_t) = \text{softmax}\big(g(f(X_t))\big),$$
$$\text{where softmax}(z)_c = \frac{e^{z_c}}{\Sigma_{c'=1}^{K} e^{z_{c'}}}. \tag{1}$$

3.1. Preliminary

We consider a dataset consisting of n i.i.d. samples, $X = \{x_i, y_i, \hat{y}_i\}_{i=1}^{n}$, where x_i is an input, y_i, and $\hat{y}_i$ are the corresponding ground truth label and pseudo-label, respectively. Following Geifman et al. (2018), we define the confidence score function $\kappa(x_i, \hat{y}_i)$ for ordinal ranking. The confidence score function $\kappa(x_i, \hat{y}_i)$ should return a high score for samples that are more likely to be correctly classified.

$$\kappa(x_i, \hat{y}_i) \leq \kappa(x_j, \hat{y}_j) \Rightarrow \Pr[\hat{y}_i = y_i] \leq \Pr[\hat{y}_j = y_j]$$
$$\text{with a high probability of } 1 - \delta, \tag{2}$$

where $0 \leq \delta \leq 1$. If $\kappa_1(\cdot)$ is a better confidence score function than $\kappa_2(\cdot)$, then $\delta_1 < \delta_2$.

The most basic and common score is Maxprob which is the maximum value of prediction $p_M(X_t)$. Another common score is negative entropy (Ent) which is a negative entropy value of prediction $p_M(X_t)$. These two scores only use $p_M(X_t)$; hence, it cannot consider the distribution of the target features. The Cossim score (Kang et al., 2019), which considers data-structure-wise knowledge, uses the cosine similarity between a sample and the center of the cluster that contains the sample based on k-means clustering. Definition of the listed scores are provided in the Appendix A.

3.2. Joint Model-Data Structure score

We propose a novel confidence score, the Joint Model-Data Structure (JMDS) score, which considers both model-wise and data-structure-wise knowledge, unlike existing confidence scores. The JMDS score consists of two confidence

scores: Log-Probability Gap (LPG) and Model Probability of Pseudo-label (MPPL) scores. Pseudo-labeling based on feature-level clustering is commonly used by other UDA methods (Kang et al., 2019; Tang et al., 2020) because the decision boundary of the model may violate the cluster assumption. In this study, GMM is used to cluster the target features and assign pseudo-labels to the target data. GMM outperforms other clustering methods in terms of confidence measurement because it provides data-structure-wise probability $p_{\text{data}}(X_t)$ (Lee et al., 2018). Details of GMM and $p_{\text{data}}(X_t)$ are provided in the Appendix B.

Data-structure-wise confidence score: We propose the Log-Probability Gap (LPG) score as a data-structure-wise confidence score because it uses the log data-structure-wise probability $\log p_{\text{data}}(X_t)$ obtained from GMM on the target feature space. First, we define MINGAP for each sample, the minimum gap from $\log p_{\text{data}}(x_i^t)_{\hat{y}_i^t}$ to the other log data-structure-wise probability value.

$$\text{MINGAP}(x_i^t) = \min_a \{\log p_{\text{data}}(x_i^t)_{\hat{y}_i^t} - \log p_{\text{data}}(x_i^t)_a\},$$

where $\hat{y}_i^t = \arg\max_c p_{\text{data}}(x_i^t)_c,\ a \in \{1, 2, \cdots, K\},\ a \neq \hat{y}_i^t$.

The LPG score is the normalized MINGAP, with a value between [0, 1].

$$\text{LPG}(x_i^t) = \frac{\text{MINGAP}(x_i^t)}{\max_j \text{MINGAP}(x_j^t)}, \tag{3}$$

where $i, j \in \{1, 2, \cdots, n_t\}$. It yields high scores for samples that is far from the decision boundary based on GMM.

Model-wise confidence score: We propose the Model Probability of Pseudo-label (MPPL) score to include knowledge of the model into the confidence score. MPPL is the model-wise probability of the corresponding pseudo-label $\hat{Y}_t$. It provides high scores for samples whose pseudo-label is the same based on $p_M(X_t)$ and $p_{\text{data}}(X_t)$.

$$\text{MPPL}(x_i^t) = p_M(x_i^t)_{\hat{y}_i^t}. \tag{4}$$

JMDS score: An overview of the JMDS score is shown in Figure 2. The JMDS score is the product of LPG and MPPL to emphasize confident samples in both scores, with a value between [0, 1]:

$$\text{JMDS}(x_i^t) = \text{LPG}(x_i^t) \cdot \text{MPPL}(x_i^t). \tag{5}$$

The JMDS score contains knowledge on the data structure from LPG and on the model from MPPL. The Maxprob and Ent scores use model prediction only, whereas the Cossim score only uses the data structure. In SFUDA, the data structure includes knowledge of the target domain and the model includes knowledge of the source domain. Therefore, the JMDS score is the only confidence score that considers knowledge from both domains. The experimental results demonstrating the superiority of the JMDS score over the other scores are presented in Section 5.1.

4. Confidence score Weighting Adaptation using the JMDS

We aim to produce high performance on the target domain by fine-tuning model M using pseudo-labels $\hat{Y}_t$ from GMM and the JMDS score. A simple way to exploit the confidence score is sample weighting which is effective for robust learning such as learning with noisy labels (Ren et al., 2018). SFUDA basically includes incorrect pseudo-labels; hence, we expect the sample weighting using confidence scores to be effective for SFUDA. Therefore, we propose a Confidence Score Weighting Adaptation using the JMDS (CoWA-JMDS) framework whose loss is as follows:

$$\mathcal{L}_{\text{CoWA-JMDS}}(x_i^t) = \text{JMDS}(x_i^t) \cdot \mathcal{L}_{\text{CE}}(p_M(x_i^t), \hat{y}_i^t), \tag{6}$$

where $\mathcal{L}_{\text{CE}}(p_M(x_i^t), \hat{y}_i^t) = -\log p_M(x_i^t)_{\hat{y}_i^t}$ is the cross entropy loss. The pseudo-code for CoWA-JMDS is provided in the Appendix D.

Weight Mixup: CoWA-JMDS rarely allows low-confidence samples whose JMDS scores close to 0 to participate in training so that the model can learn robustly with incorrect pseudo-labels. However, this means that the knowledge provided by the target feature distribution is not fully utilized. Therefore, we propose a technique called weight Mixup, a variant of Mixup (Zhang et al., 2017), to utilize more knowledge of the target feature distribution.

Mixup mixes images and corresponding labels. All mixed samples had the same sample-wise weights for training. However, in CoWA-JMDS, the sample has its own confidence score as a sample-wise weight for training. This means that mixed images that use low-confidence samples for mixing should have a lower sample-wise weight for Mixup training. Therefore, the proposed weight Mixup mixes the corresponding sample-wise weights together.

$$\begin{aligned} \tilde{x}^t &= \gamma \cdot x_i^t + (1-\gamma) \cdot x_j^t, \\ \tilde{y}^t &= \gamma \cdot o(\hat{y}_i^t) + (1-\gamma) \cdot o(\hat{y}_j^t), \\ w(\tilde{x}^t) &= \gamma \cdot \text{JMDS}(x_i^t) + (1-\gamma) \cdot \text{JMDS}(x_j^t), \end{aligned} \tag{7}$$

where $\gamma \sim \text{Beta}(\alpha, \alpha)$, for $\alpha \in (0, \infty)$, and $o(\cdot)$ is a one-hot encoding function. The loss function of the weight Mixup is modified as follows:

$$\mathcal{L}_{\text{Mixup}}(\tilde{x}^t, \tilde{y}^t) = w(\tilde{x}^t) \cdot \mathbb{E}_{\tilde{y}^t}[-\log p_M(\tilde{x}^t)] \tag{8}$$

Weight Mixup makes the training robust to incorrect pseudo-labels in the following manner: A mixture of low- and high-confidence samples will produce a sample with mid-level

Table 1. Evaluation of the JMDS score based on AURC.

Dataset	Task	Naïve PL+Maxprob	Naïve PL+Ent	SSPL+Cossim	GMM+Cossim	GMM+MPPL	GMM+LPG	GMM+JMDS
Office-31	A → D	0.047	0.051	**0.018**	0.031	0.039	0.033	0.033
	A → W	0.074	0.081	**0.034**	0.045	0.059	0.042	0.044
	D → A	0.158	0.165	0.140	0.130	0.131	0.127	**0.115**
	D → W	0.007	0.008	0.009	0.009	0.005	0.004	**0.004**
	W → A	0.157	0.167	**0.107**	0.108	0.132	0.120	0.113
	W → D	0.002	0.002	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
	Avg.	0.074	0.079	**0.052**	0.054	0.061	0.055	**0.052**
Office-Home	Ar → Cl	0.308	0.316	0.296	0.274	0.278	0.265	**0.256**
	Ar → Pr	0.140	0.145	**0.100**	0.105	0.116	0.125	0.104
	Ar → Rw	0.088	0.095	0.086	0.086	0.076	0.086	**0.068**
	Cl → Ar	0.238	0.249	0.200	0.194	0.212	0.216	**0.197**
	Cl → Pr	0.159	0.168	**0.105**	0.113	0.131	0.125	0.115
	Cl → Rw	0.151	0.159	0.113	0.113	0.125	0.115	**0.106**
	Pr → Ar	0.237	0.246	0.185	**0.184**	0.210	0.214	0.190
	Pr → Cl	0.365	0.375	0.339	0.315	0.327	0.293	**0.293**
	Pr → Rw	0.095	0.099	0.080	0.082	0.084	0.091	**0.073**
	Rw → Ar	0.138	0.147	0.129	0.125	0.126	0.154	**0.118**
	Rw → Cl	0.314	0.325	0.298	0.284	0.275	0.248	**0.238**
	Rw → Pr	0.073	0.078	0.062	0.063	0.065	0.078	**0.059**
	Avg.	0.192	0.200	0.166	0.162	0.169	0.168	**0.151**
VisDA-2017	T → V	0.274	0.284	0.261	0.202	0.204	0.172	**0.162**

confidence, which can robustly and effectively participate in the learning. By contrast, the resulting sample will be suppressed when low-confidence samples are mixed with each other. Therefore, weight Mixup is helpful for CoWA-JMDS as it enhances target feature knowledge by augmenting samples with the mid-level confidence.

4.1. General UDA scenarios

CoWA-JMDS can be easily extended to more general UDA settings. Open-set (Panareda Busto & Gall, 2017) and partial-set (Cao et al., 2018) scenarios are relatively realistic scenarios where only a few categories of interest are shared between the source and target data.

In an open-set scenario, the target domain contains unseen classes that are not included in the source domain. Therefore, we should classify known classes included in the source domain and unknown classes. Algorithm 1 in the Appendix C shows the process of classifying the known and unknown classes. Following the protocols in Liang et al. (2020a), we sorted the entropy of the samples and performed a two-class k-means clustering. Then, a high-entropy cluster was classified as unknown samples, whereas a low-entropy cluster was classified as known samples. Known samples were used to train the models.

In a partial-set scenario, the target domain contains a few classes included in the source domain. Therefore, the absent classes should be filtered out. Algorithm 2 in the Appendix C shows how to estimate the classes included in the target domain. First, we initialized the parameters for GMM using the prediction of the model and perform the GMM expectation-maximization iteration once. Next, we filtered out classes whose number of samples were lower than a threshold. Finally, the aforementioned two steps were iteratively executed until there were no filtered out classes.

Table 2. Accuracy (%) on Office-31 dataset for UDA and SFUDA methods (ResNet-50).

Task	Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
SFUDA	SFIT (Hou & Zheng, 2021)	89.9	91.8	73.9	98.7	72.0	99.9	87.7
	SHOT (Liang et al., 2020a)	94.0	90.1	74.7	98.4	74.3	99.9	88.6
	3C-GAN (Li et al., 2020)	92.7	93.7	75.3	98.5	**77.8**	99.8	89.6
	NRC (Yang et al., 2021)	**96.0**	90.8	75.3	**99.0**	75.0	**100.0**	89.4
	CoWA-JMDS (w/o weight Mixup)	93.7	93.5	75.5	98.0	76.8	99.8	89.6
	CoWA-JMDS	94.4	**95.2**	**76.2**	98.5	77.6	99.8	**90.3**
UDA	ResNet (He et al., 2016)	68.9	68.4	62.5	96.7	60.7	99.3	76.1
	CAN (Kang et al., 2019)	95.0	94.5	78.0	99.1	77.0	99.8	90.6
	RSDA-MSTN (Gu et al., 2020)	95.8	96.1	77.4	99.3	78.9	100	91.1
	FixBi (Na et al., 2020)	95.0	96.1	78.7	99.3	79.4	100	91.4

5. Experiments

We evaluated our proposed methods, the JMDS score and CoWA-JMDS, on three public UDA benchmarks: Office-31 (Saenko et al., 2010), Office-Home (Venkateswara et al., 2017), and VisDA-2017 (Peng et al., 2017). More details are provided in the Appendix E.

5.1. JMDS evaluation

We introduced the JMDS score as a reliable confidence score. To prove its efficacy, we compared the JMDS score with other scores (Mandelbaum & Weinshall, 2017; Kang et al., 2019). The pseudo-label of Maxprob and Ent scores is given by the index of maximum model probability $\text{argmax}_c\, p_M(x_i^t)_c$ following naive PL (Lee et al., 2013). SSPL, proposed by Liang et al. (2020a), and GMM provide the pseudo-label for Cossim. SSPL uses modified k-means clustering to generate pseudo-labels.

To evaluate the confidence score, we measured the Area Under Risk-Coverage curve (AURC) (Ding et al., 2020)

Table 3. Accuracy (%) on Office-Home for UDA and source-free UDA methods (ResNet-50).

Task	Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
SFUDA	BAIT (Yang et al., 2020)	57.4	77.5	**82.4**	68.0	77.2	75.1	67.1	55.5	81.9	**73.9**	59.5	84.2	71.6
	SHOT (Liang et al., 2020a)	57.1	78.1	81.5	68.0	78.2	78.1	67.4	54.9	82.2	73.3	58.8	84.3	71.8
	NRC (Yang et al., 2021)	**57.7**	**80.3**	82.0	68.1	79.8	78.6	65.3	56.4	**83.0**	71.0	58.6	**85.6**	72.2
	CoWA-JMDS (w/o weight Mixup)	56.4	78.6	80.3	68.8	79.7	78.7	**68.1**	56.8	82.0	73.4	59.1	83.9	72.2
	CoWA-JMDS	56.9	78.4	81.0	**69.1**	**80.0**	**79.9**	67.7	**57.2**	82.4	72.8	**60.5**	84.5	**72.5**
UDA	ResNet-50 (He et al., 2016)	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
	RSDA-MSTN (Gu et al., 2020)	53.2	77.7	81.3	66.4	74.0	76.5	67.9	53.0	82.0	75.8	57.8	85.4	70.9
	FixBi (Na et al., 2020)	58.1	77.3	80.4	67.7	79.5	78.1	65.8	57.9	81.7	76.4	62.9	86.7	72.7

Table 4. Accuracy (%) on VisDA-2017 for UDA and source-free UDA methods (ResNet-101).

Task	Method	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Average
SFUDA	SFIT (Hou & Zheng, 2021)	94.3	79.0	**84.9**	63.6	92.6	92.0	88.4	79.1	92.2	79.8	87.6	43.0	81.4
	3C-GAN (Li et al., 2020)	94.8	73.4	68.8	**74.8**	93.1	95.4	88.6	84.7	89.1	84.7	83.5	48.1	81.6
	SHOT (Liang et al., 2020a)	94.3	88.5	80.1	57.3	93.1	94.9	80.7	80.3	91.5	89.1	86.3	58.2	82.9
	NRC (Yang et al., 2021)	**96.8**	**91.3**	82.4	62.4	96.2	95.9	86.1	80.6	**94.8**	**94.1**	**90.4**	**59.7**	85.9
	CoWA-JMDS (w/o weight Mixup)	96.3	88.5	84.1	59.7	95.2	96.9	82.1	82.3	93.3	92.8	87.5	51.1	84.2
	CoWA-JMDS	96.2	89.7	**83.9**	73.8	**96.4**	**97.4**	**89.3**	**86.8**	94.6	92.1	88.7	53.8	**86.9**
UDA	ResNet-101 (He et al., 2016)	72.3	6.1	63.4	91.7	52.7	7.9	80.1	5.6	90.1	18.5	78.1	25.9	49.4
	CAN (Kang et al., 2019)	97.0	87.2	82.5	74.3	97.8	96.2	90.8	80.7	96.6	96.3	87.5	59.9	87.2
	FixBi (Na et al., 2020)	96.1	87.8	90.5	90.3	96.8	95.3	92.8	88.7	97.2	94.2	90.9	25.7	87.2

using the 0/1 loss function, which returns a value of one if the pseudo- and ground truth-labels of the sample are different, and a value of zero if they are the same. Ding et al. (2020) compared the Area Under Receiver Operating Characteristic curve (AUROC), Area Under Precision-Recall curve (AUPR), and AURC. They claimed that AURC is the only reliable measurement when the underlying models are the same. The risk-coverage curve was first proposed by Geifman et al. (2018). After obtaining the high-confidence set, $X_t^h = \{x_i^t | \kappa(x_i^t, \hat{y}_i^t) > \tau\}$, where τ is a threshold, risk is the average empirical loss of X_t^h, and the coverage is $|X_t^h|/|X_t|$. A lower AURC value indicates higher reliability because it implies a lower risk for the same coverage. When 0/1 loss is applied, a high AURC indicates low correctness for corresponding pseudo-labels.

The experimental results are presented in Table 1. We measured the reliability of various confidence scores for the pre-trained source model obtained using five random seeds. The best AURC is indicated in bold. We quantitatively compared various confidence measuring strategies based on AURC values. The JMDS score achieved the lowest AURC value in most adaptation tasks. Using the MPPL or LPG score alone is worse than using the JMDS score, which considers both scores. This demonstrates that using knowledge of the model and data structure jointly is superior to considering only one aspect.

5.2. CoWA-JMDS evaluation

We evaluated model M trained by CoWA-JMDS and compared it with SFUDA (Liang et al., 2020a; Li et al., 2020; Yang et al., 2020; Hou & Zheng, 2021; Yang et al., 2021) and conventional UDA baseline methods (Kang et al., 2019; Gu et al., 2020; Na et al., 2020). Notably, our task is SFUDA, which is a more challenging task than conventional UDA that directly uses the source data during training. We trained the models using five different random seeds and reported their average performance. The best accuracy is indicated in bold, and the second-best accuracy is underlined.

Closed-set scenario: Table 2, 3, and 4 show the classification accuracy for a closed-set scenario in all tasks on each dataset: Office-31, Office-Home, and VisDA-2017.

CoWA-JMDS achieved the best performance for all three datasets. CoWA-JMDS boosts the best performance 0.7% on the Office-31 dataset, 0.3% on the Office-Home dataset, and 1.0% on the VisDA-2017 dataset. The results demonstrate that our proposed CoWA-JMDS framework is effective for SFUDA. Even, despite its simplicity, we obtained the same performance with the state-of-the-art method on the Office-31, and Office-Home dataset without weight Mixup, only using the JMDS score as a sample weight and the cross entropy loss.

Extended UDA scenarios: We evaluated CoWA-JMDS for open-set and partial-set scenarios on the Office-Home dataset. Following the protocols in Liang et al. (2020a), the source domain consists of 25 classes (the first 25 in alphabet order) but the target domain contains 65 classes including unknown samples for an open-set scenario. However, for a partial-set scenario, the source domain consists of 65 classes but the target domain contains the same 25 classes.

Table 5 shows the results of experiments for the open-set and partial-set scenarios. CoWA-JMDS achieved the best performance among all methods including conventional UDA methods for both scenarios. In an open-set scenario, CoWA-JMDS without weight Mixup outperforms CoWA-JMDS with weight Mixup because weight Mixup had a negative effect when the known-classified unknown class samples were mixed with other samples.

Table 5. Accuracy (%) on Office-Home for open-set and partial-set scenarios (ResNet-50).

Task (Open-set)	Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
SFUDA	SHOT (Liang et al., 2020a)	64.5	**80.4**	84.7	63.1	75.4	81.2	65.3	**59.3**	83.3	69.6	**64.6**	82.3	72.8
	CoWA-JMDS (w/o weight Mixup)	**64.6**	80.2	**88.1**	67.3	83.5	**82.2**	63.9	57.1	**84.4**	**70.8**	64.0	84.8	**74.2**
	CoWA-JMDS	63.3	79.2	85.4	**67.6**	**83.6**	82.0	**66.9**	56.9	81.1	68.5	57.9	**85.9**	73.2
UDA	ResNet-50 (He et al., 2016)	53.4	52.7	51.9	69.3	61.8	74.1	61.4	64.0	70.0	78.7	71.0	74.9	64.3
	STA (Liu et al., 2019)	58.1	53.1	54.4	71.6	69.3	81.9	63.4	65.2	74.9	85.0	75.8	80.8	69.5
	PGL (Luo et al., 2020)	61.6	77.1	85.9	68.8	72.0	82.8	72.2	58.4	82.6	78.6	65.0	83.0	74.0
Task (Partial-set)	Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg
SFUDA	SHOT (Liang et al., 2020a)	64.8	85.2	**92.7**	**76.3**	77.6	88.8	79.7	64.3	89.5	80.6	66.4	85.8	79.3
	CoWA-JMDS (w/o weight Mixup)	**69.7**	91.6	92.1	**78.9**	**86.3**	91.6	**81.5**	64.4	89.7	**84.1**	71.6	90.2	82.6
	CoWA-JMDS	69.6	**93.2**	92.3	**78.9**	81.3	**92.1**	79.8	**71.7**	**90.0**	83.8	**72.2**	**93.7**	**83.2**
UDA	ResNet-50 (He et al., 2016)	46.3	67.5	75.9	59.1	59.9	62.7	58.2	41.8	74.9	67.4	48.2	74.2	61.3
	SAFN (Xu et al., 2019)	58.9	76.3	81.4	70.4	73.0	77.8	72.4	55.3	80.4	75.8	60.4	79.9	71.8
	BA^{3}US (Liang et al., 2020b)	60.6	83.1	88.4	71.8	72.8	83.4	75.5	61.6	86.5	79.3	62.8	86.1	76.0

6. Further analysis

In this section, we analyze how the proposed components, the JMDS score and weight Mixup, work through additional experimental results.

JMDS score at the pre-trained source model: We considered a toy example to describe how the JMDS score works in Figure 3. Figure 3(a) and 3(f) show the source and target data distributions, respectively. We generated three-mode Gaussian source features and trained a two-layer fully-connected classifier. Then, we generated another three-mode Gaussian target feature with a slightly different distribution than the source features and tested how various confidence scores are distributed. Figure 3(b), 3(g)-3(j) follow the decision boundary of the model, and Figure 3(c)-3(e) follow the decision boundary of GMM.

The model-wise confidence score, Maxprob and Ent scores, have almost the same distribution as shown in Figure 3(b) and 3(g). Because these two scores used the decision boundary of the model, as shown in Figure 3(f), there were many confidently misclassified samples. In Figure 3(h), when using MPPL scores, confusing samples, which are in an overlapped region have lower confidence scores than the aforementioned two scores. However, MPPL scores still give overconfident scores for most of the samples, because DNNs using ReLU are overconfident for most of the data (Hein et al., 2019). It causes negative effects on training.

In Figure 3(d), LPG scores work well for the decision boundary of GMM. However, in Figure 3(i), samples near the decision boundary of the model in class 3 have high confidence because they do not use model-wise knowledge. This is undesirable to confidence scores because samples near the decision boundary of the model should be suppressed. Based on the cluster assumption (Grandvalet et al., 2005), the decision boundary of the model does not pass through the training data when they fit. In other words, features near the decision boundary of the model are far from the training feature distribution, such as out-of-distribution samples. Therefore, the model cannot provide precise predictions for these features; hence, they need to be suppressed.

By considering both model-wise and data-structure-wise knowledge, the JMDS score, which combines MPPL and LPG scores, can overcome the aforementioned problems of overconfidence and out-of-distribution problems. As shown in Figure 3(e) and 3(j), the JMDS score shows a reliable distribution in terms of both model-wise and data-structure-wise knowledge. Quantitatively, the JMDS score achieved the best AURC value in SFUDA as shown in Table 1. Qualitatively, as shown in Figure 4(a), the JMDS score has the lowest risk for any coverage at the pre-trained source model on the VisDA-2017 dataset.

JMDS score during learning: Figure 3 and Table 1 show that the JMDS score is a reasonable confidence score in the epoch 0 state, that is, without learning. It is necessary to check whether it is effective when it is applied as a sample-wise weight during learning. According to Ren et al. (2018), one crucial advantage of sample weighting is robustness against training set bias such as label noise. Therefore, CoWA-JMDS shows robustness to incorrect pseudo-labels through sample weighting using the JMDS score so that it performs well in SFUDA.

As shown in Figure 4(b), sample weighting using confidence scores (LPG, MPPL, and JMDS) improved the performance compared to when sample weighting was not performed (None). Additionally, sample weighting using JMDS was better than when sample weighting was performed using MPPL or LPG alone. This proves that the JMDS score has a positive effect not only on the pre-trained source model but also on the actual SFUDA learning.

Figure 3(e) shows many samples had low JMDS values. Low values can cause an underfitting problem during learning. However, as CoWA-JMDS training progressed, the JMDS scores also increased, shown in Figure 4(c). It indicates that the participation of data in learning increases as learning continues. This prevents the underfitting problem and allows CoWA-JMDS to learn properly using all samples without sample selection. The result is related to a progres-

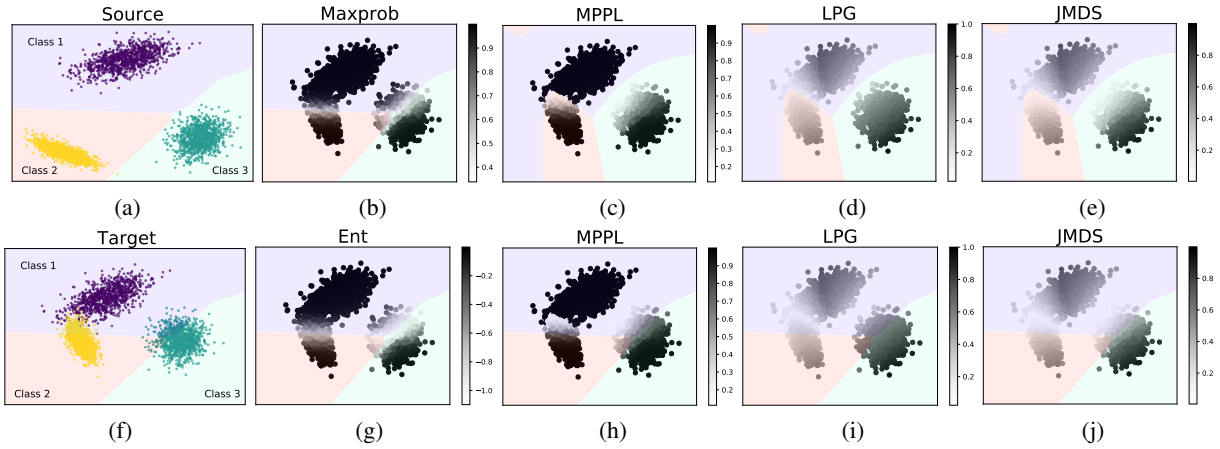


Figure 3. A toy example to determine how the JMDS score works. (a) The source feature distribution. (f) The target feature distribution. (b)-(e), (g)-(j) The higher the confidence score, the darker the point. (c)-(e) use the decision boundary of GMM and the others use the decision boundary of the pre-trained source model.

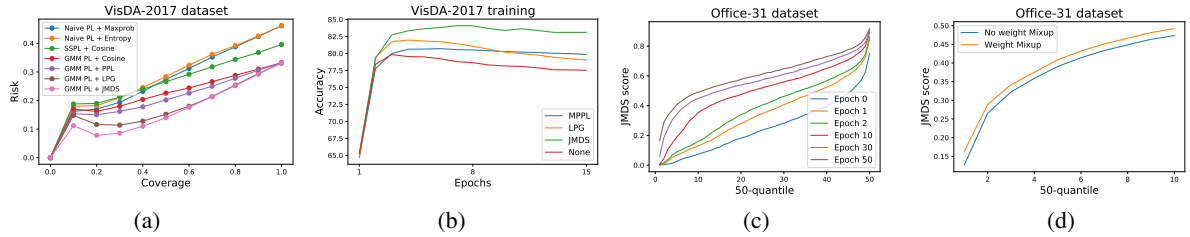


Figure 4. (a) The RC curve of the pre-trained source model on the VisDA-2017 dataset. (b) The accuracy-epoch plot on the VisDA-2017 dataset. (c) The JMDS-quantile plot on various epochs. (d) The JMDS-quantile plot to demonstrate the effectiveness of weight Mixup.

sive learning scheme, such as curriculum learning (Bengio et al., 2009) and self-paced learning (Kumar et al., 2010).

Effect of weight Mixup: We proposed weight Mixup to exploit more knowledge of the target feature distribution with low-confidence samples during CoWA-JMDS. Figure 4(d) shows the effect of weight Mixup during CoWA-JMDS training. Because low-confidence samples are indirectly more included in learning as a form of mixed mid-level confidence samples, weight Mixup boosts the JMDS score of low-confidence samples. It encourages the model can utilize more knowledge of the target feature distribution than when weight Mixup is not used.

As shown in Table 6, naïve use of Mixup leads marginal improvement in terms of average accuracy. Mixup boosts 0.2% and 0.3% for SHOT and the pseudo-labels of GMM, respectively. This is still due to the vulnerability of the confirmation bias problem because Mixup gives the same importance to the mixed samples without considering the incorrect pseudo-labels in SFUDA. Therefore, we need to use weight Mixup in SFUDA. Weight Mixup increases performance by 3.4% compared to Mixup with the pseudo-labels of GMM which utilizes the same pseudo-labels. Also, weight Mixup boosts 0.7% when it is applied to CoWA-JMDS. It demonstrates that weight Mixup is an effective technique that can be used in learning with sample weighting based on confidence scores.

Table 6. A weight mixup experiment on Office-31 dataset.

Method	Avg.
SHOT (Liang et al., 2020a)	88.6
SHOT + Mixup (Zhang et al., 2017)	88.8
GMM PL	86.6
GMM PL + Mixup	86.9
GMM PL + JMDS	89.6
GMM PL + JMDS + weight Mixup (CoWA-JMDS)	90.3

7. Conclusion

In this study, we propose the JMDS score, the novel confidence score for SFUDA to differentiate sample importance based on confidence for pseudo-labels. The JMDS score jointly considers both of the source and target domain knowledge unlike existing scores. We then propose CoWA-JMDS, the SFUDA method that uses the JMDS score as a sample-wise weight. CoWA-JMDS also uses weight Mixup, which is the proposed variant of Mixup, to exploit more target domain knowledge. CoWA-JMDS can be extended to more realistic scenarios, such as open-set and partial-set scenarios. The experiments showed that the proposed JMDS score and CoWA-JMDS achieve state-of-the-art performance on UDA benchmarks in various SFUDA scenarios.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A3B1077720), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)], Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2022-0-00959), AIRS Company in Hyundai Motor and Kia through HMC/KIA-SNU AI Consortium Fund, Samsung Electronics Co., Ltd. (Mobile Communications Business), and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2022.

References

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Cao, Z., Long, M., Wang, J., and Jordan, M. I. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2724–2732, 2018.

Ding, Y., Liu, J., Xiong, J., and Shi, Y. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4–5, 2020.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. *arXiv preprint arXiv:1705.08500*, 2017.

Geifman, Y., Uziel, G., and El-Yaniv, R. Bias-reduced uncertainty estimation for deep neural classifiers. *arXiv preprint arXiv:1805.08206*, 2018.

Grandvalet, Y., Bengio, Y., et al. Semi-supervised learning by entropy minimization. *CAP*, 367:281–296, 2005.

Gu, X., Sun, J., and Xu, Z. Spherical space domain adaptation with robust pseudo-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9101–9110, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hein, M., Andriushchenko, M., and Bitterwolf, J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41–50, 2019.

Hou, Y. and Zheng, L. Visualizing adapted knowledge in domain transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13824–13833, 2021.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR, 2015.

Kang, G., Jiang, L., Yang, Y., and Hauptmann, A. G. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4893–4902, 2019.

Kumar, M. P., Packer, B., and Koller, D. Self-paced learning for latent variable models. In *NIPS*, volume 1, pp. 2, 2010.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.

Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Li, R., Jiao, Q., Cao, W., Wong, H.-S., and Wu, S. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9641–9650, 2020.

Liang, J., Hu, D., and Feng, J. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 6028–6039. PMLR, 2020a.

Liang, J., Wang, Y., Hu, D., He, R., and Feng, J. A balanced and uncertainty-aware approach for partial domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 123–140. Springer, 2020b.

Liu, H., Cao, Z., Long, M., Wang, J., and Yang, Q. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2927–2936, 2019.

Luo, Y., Wang, Z., Huang, Z., and Baktashmotlagh, M. Progressive graph learning for open-set domain adaptation. In *International Conference on Machine Learning*, pp. 6468–6478. PMLR, 2020.

Mandelbaum, A. and Weinshall, D. Distance-based confidence score for neural network classifiers. *arXiv preprint arXiv:1709.09844*, 2017.

Müller, R., Kornblith, S., and Hinton, G. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019.

Na, J., Jung, H., Chang, H., and Hwang, W. Fixbi: Bridging domain spaces for unsupervised domain adaptation. *arXiv preprint arXiv:2011.09230*, 2020.

Nair, T., Precup, D., Arnold, D. L., and Arbel, T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2009.

Panareda Busto, P. and Gall, J. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 754–763, 2017.

Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., and Saenko, K. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pp. 4334–4343. PMLR, 2018.

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

Salimans, T. and Kingma, D. P. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *arXiv preprint arXiv:1602.07868*, 2016.

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

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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

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

Yang, S., Wang, Y., van de Weijer, J., Herranz, L., and Jui, S. Unsupervised domain adaptation without source data by casting a bait. *arXiv preprint arXiv:2010.12427*, 2020.

Yang, S., van de Weijer, J., Herranz, L., Jui, S., et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

A. Previous confidence scores

We redefine the scores to be between [0,1].

$$\text{Maxprob}(x_i^t) = \max_c p_M(x_i^t)_c,$$
$$\text{Ent}(x_i^t) = 1 + \frac{\Sigma_{c=1}^K p_M(x_i^t)_c \log p_M(x_i^t)_c}{\log K},$$
$$\text{Cossim}(x_i^t) = \frac{1}{2}(1 + \frac{\langle x_i^t, C_{\hat{y}_i^t}\rangle}{\|x_i^t\|\|C_{\hat{y}_i^t}\|}),$$
$$\text{where } p(x)_c = p(\hat{y} = c|x),$$
$$C_{\hat{y}_i^t} \text{ is the center of the cluster corresponding to class } \hat{y}_i^t.$$

B. Gaussian Mixture Modeling (GMM)

Using GMM, we obtain the parameters μ_c, Σ_c, and π_c and log-likelihood

$$\begin{aligned}\log p(x_i^t|\mu_c, \Sigma_c) = -\frac{1}{2}(d \log 2\pi + \log|\Sigma_c| + \\ (f(x_i^t) - \mu_c)^T \Sigma_c^{-1}(f(x_i^t) - \mu_c)),\end{aligned} \tag{9}$$

where π_c, μ_c and Σ_c are the mixing coefficient, mean vector and covariance matrix of the class $c \in \{1, 2, \cdots, K\}$ respectively. Now, we obtain the data-structure-wise probability $p_{\text{data}}(x_i^t)_c$ for class c based on the ratio of log-likelihood and the corresponding pseudo-label:

$$\begin{aligned}p_{\text{data}}(x_i^t)_c &= \frac{\pi_c(x_i^t)p(x_i^t|\mu_c, \Sigma_c)}{\Sigma_{c'}\{\pi_{c'}(x_i^t)p(x_i^t|\mu_{c'}, \Sigma_{c'})\}} \\ \text{where} \quad & c, c' \in \{1, 2, \cdots, K\}. \\ \hat{y_i} &= \arg\max_c p_{\text{data}}(x_i^t)_c\end{aligned} \tag{10}$$

C. Algorithms

Algorithm 1 shows the full process for known/unknown classification for an open-set scenario. Algorithm 2 shows the full process for class estimation for a partial-set scenario.

Algorithm 1 Known/unknown classification

1: **Input:** Unlabeled target data X_t, the model $M = g \circ f$.
2: Compute $p_M(X_t) = \text{softmax}\big(g(f(X_t))\big)$.
3: Compute the entropy of $p_M(X_t)$.
4: Divide X_t in the low-entropy cluster $C_{\text{low-entropy}}$ and the high-entropy cluster $C_{\text{high-entropy}}$ based on 2-class k-means.
5: **Output:** $C_{\text{low-entropy}}$.

Algorithm 2 Class estimation

Input: Unlabeled target data X_t, the model $M = g \circ f$, and a threshold τ.
$C_{\text{partial}} = \{c_1, \ldots, c_K\}$.
repeat
  Initialize π, μ, Σ based on $p_M(X_t)$ and C_{partial}.
  Perform one EM iteration for GMM.
  compute $p_{\text{data}}(X_t)$
  for $i = 1$ **to** $|C_{\text{partial}}|$ **do**
    if $\sum_{i=1}^{n_t} p_{\text{data}}(x_i)_{c_j} < \tau \cdot \frac{n_t}{|C_{\text{partial}}|}$ **then**
      $C_{\text{partial}} \leftarrow C_{\text{partial}} \backslash c_j$
    end if
  end for
until C_{partial} converges
Output: C_{partial}

D. Pseudo-code for CoWA-JMDS

Algorithm 3 shows the full procedure of CoWA-JMDS. Full code is available at a supplementary file.

Algorithm 3 CoWA-JMDS

Input: Unlabeled target data X_t, the model $M = g \circ f$.
epoch $\leftarrow 0$.
repeat
  if Partial-set scenario **then**
    Perform class estimation.
  end if
  Perform GMM on $f(X_t)$ and compute $p_{\text{data}}(X_t)$.
  if Open-set scenario **then**
    Perform known/unknown classification.
  end if
  Compute JMDS score using Equation (5).
  for $i \leftarrow 1$ **to** $iterations\_per\_epoch$ **do**
    if No weight Mixup **then**
      Compute loss using Equation (6).
    else if Weight Mixup **then**
      Obtain mixed inputs $\tilde{x}^t$, pseudo-labels $\tilde{y}^t$, and JMDS scores $\tilde{w}^t$ using Equation (7).
      Compute loss using Equation (8).
    end if
    Update the model M using loss.
  end for
  epoch $\leftarrow$ epoch$+1$.
until epoch $< max\_epoch$

E. Implementation details

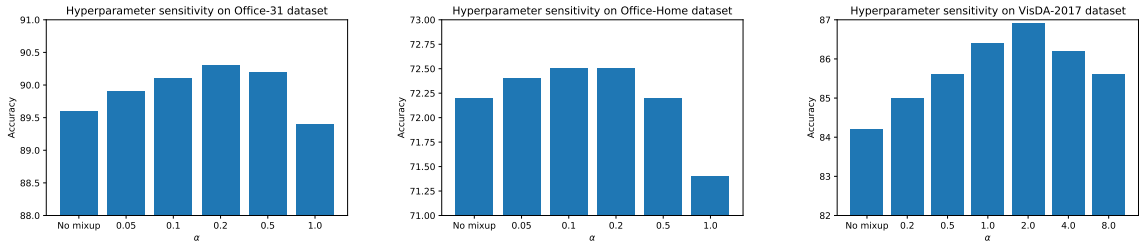
Office-31 (Saenko et al., 2010) is a small-sized standard UDA benchmark with three domains from different sources, that is, collected from the Amazon website (A), Web camera (W), and DSLR (D). The dataset contain 4,110 images of 31 object classes of office supplies. Office-Home (Venkateswara et al., 2017) is a medium-sized UDA benchmark that contains Artistic images (Ar), Clip Art (Cl), Product images (Pr), and Real-World images (Rw). The dataset contain 15,500 images of 65 object classes. VISDA-2017 (Peng et al., 2017) is a challenging large-sized UDA benchmark. It contains a training dataset with 152,397 synthetic data and a test dataset with 55,388 real images with 12 categories.

We use ResNet-50 or ResNet-101 (He et al., 2016) pre-trained on the source data as our backbone network. Following

Liang et al.(Liang et al., 2020a), we put a bottleneck layer with 256 units and a task-specific classifier layer. a weight normalization layer (Salimans & Kingma, 2016) is applied in last classifier layer and bottleneck layer consists batch normalization layer (Ioffe & Szegedy, 2015). Pre-trained source model trained with label smoothing (Müller et al., 2019). We use mini-batch SGD with momentum 0.9 and weight decay 1e-3 for all experiments. Learning rate of a bottleneck layer is 1e-2 and the remainders are 1e-3 for all three datasets: the Office-31, Office-Home, VisDA-2017 datasets. We do not use learning rate decay and the number of epochs are 50, 30, and 15, respectively. We set batch size 64 for all three benchmark datasets. The hyperparameter of weight Mixup α is set to 0.2, 0.2, and 2.0, respectively. The threshold τ for class estimation in a partial-set scenario is set to 0.3 for the Office-Home dataset. Since we perform GMM in the high-dimensional feature space, Expectation-Maximization iteration is conducted once to resolve the instability.

F. Hyperparameter sensitivity

Weight mixup has a hyperparameter α which decides the mixing coefficient γ. Figure 5 shows the final accuracy of CoWA-JMDS for various values of α.



(a) (b) (c)

Figure 5. Hyperparameter sensitivity result of α on three datasets.

In a partial-set scenario, there is a hyperparameter τ to estimate which classes are included in the target domain. Figure 6 shows the accuracy of CoWA-JMDS for various values of τ.

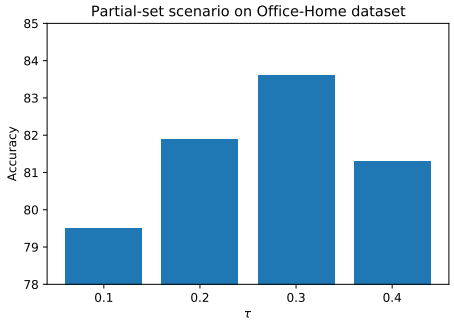


Figure 6. Hyperparameter sensitivity result of τ on the Office-Home dataset.