Balanced Open Set Domain Adaptation via Centroid Alignment

Mengmeng Jing[1], **Jingjing Li**[1*], **Lei Zhu**[2], **Zhengming Ding**[3], **Ke Lu**[1], **Yang Yang** [1]
[1] University of Electronic Science and Technology of China
[2] Shandong Normal University,
[3] Department of Computer Science, Tulane University
jingmeng1992@gmail.com, lijin117@yeah.net, leizhu0608@gmail.com, zding1@tulane.edu, kel@uestc.edu.cn, dlyyang@gmail.com

Abstract

Open Set Domain Adaptation (OSDA) is a challenging domain adaptation setting which allows the existence of unknown classes on the target domain. Although existing OSDA methods are good at classifying samples of known classes, they ignore the classification ability for the unknown samples, making them unbalanced OSDA methods. To alleviate this problem, we propose a balanced OSDA methods which could recognize the unknown samples while maintain high classification performance for the known samples. Specifically, to reduce the domain gaps, we first project the features to a hyperspherical latent space. In this space, we propose to bound the centroid deviation angles to not only increase the intra-class compactness but also enlarge the inter-class margins. With the bounded centroid deviation angles, we employ the statistical Extreme Value Theory to recognize the unknown samples that are misclassified into known classes. In addition, to learn better centroids, we propose an improved centroid update strategy based on sample reweighting and adaptive update rate to cooperate with centroid alignment. Experimental results on three OSDA benchmarks verify that our method can significantly outperform the compared methods and reduce the proportion of the unknown samples being misclassified into known classes.

Introduction

In the field of artificial intelligence, it is common to make cross-domain knowledge transfer, e.g., cross-domain semantic segmentation (Hoffman et al. 2018; Saito et al. 2018a), cross-domain object classification (Long et al. 2018b; Jing et al. 2020; Li et al. 2018b,a, 2019, 2020) and cross-language text classification (Prettenhofer and Stein 2010). However, in real-world applications, a more common situation is that the target domain is mixed with samples of unknown classes. Traditional domain adaptation methods would fail to cope with this situation. Therefore, Open Set Domain Adaptation (OSDA) is proposed to address the problem of domain adaptation with unknown classes in the target domain (Baktashmotlagh et al. 2018; Panareda Busto and Gall 2017).

There are two major challenges in OSDA: (1) The large inter-domain gaps make the well-trained classifier in the source domain suffers a large classification risk in the target domain. (2) The existence of the unknown samples makes it difficult to recognize the unknown samples while maintain the classification performance for the known samples. Existing methods generally handle OSDA tasks by addressing the above two problems. For example, Saito et al. (Saito et al. 2018b) train a classifier to build a decision boundary for known and unknown, and train a generator to make target samples far from the boundary. Liu et al. (Liu et al. 2019) adopt a coarse-to-fine weighting mechanism to progressively separate the samples of known and unknown.

Although many OSDA methods are good at classifying the known samples, they ignore the classification ability for the unknown samples, making them unbalanced OSDA methods. Specifically, there are two widely used evaluation metrics for OSDA, i.e., the mean accuracy for all classes (OS_{acc}) and the mean accuracy for the known classes (OS^*_{acc}) (Panareda Busto and Gall 2017). Under these metrics, many state-of-the-art methods achieve high OS_{acc} and OS^*_{acc}, but in fact, they have low classification performance for unknown samples. For example, in Office-31 dataset, STA (Liu et al. 2019) has a high average accuracy for samples of the known classes (OS^*_{acc}=94.6%), but its accuracy for unknown class samples is only 50.5% (see Table 1). Since all unknown classes are regarded as one class, with the increase of the number of known classes, the smaller the contribution of unknown samples to the total OS_{acc}, the easier they are to be ignored by existing methods. Low classification performance for unknown samples causes a high false positive rate for known classes, which reduces the practical value of many OSDA methods. In addition, ignoring the classification ability for the unknown samples results in that many existing OSDA methods suffer from high open set risk (Luo et al. 2020; Fang et al. 2020).

Considering the unbalanced problem, in this paper, we argue that we should pay equal attention to the classification performance of known and unknown classes. To this end, we propose a balanced OSDA method based on centroid alignment in a hyperspherical latent space. The "balanced" means that our method can recognize the unknown class samples while maintain high classification performance for the known class samples. The motivation of our method is that, we take the centroids of the classes in the source domain as the centroids of samples in both domains, and use the centroid deviation angle to measure the discrepancy be-

*Corresponding Author.

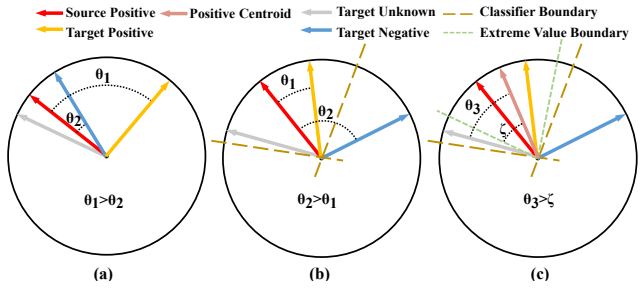


Figure 1: (a) Samples randomly distributed on the hypersphere where each sample is a vector and the intra-class angle θ_1 may be larger than the inter-class angle θ_2. (b) The centroid deviation angles are bounded so that the intra-class angle θ_1 is minimized while the inter-class angle θ_2 is maximized. Meanwhile, an unknown sample is misclassified into a known class. (c) The unknown sample is rejected as unknown due to its extremely larger centroid deviation angle.

tween a sample and a centroid. Then, we bound the centroid deviation angles to not only increase margins of the inter-class samples but also enhance the compactness of the intra-class samples, as well as reduce domain gaps. With the bounded centroid deviation angles, we further employ the statistical Extreme Value Theory (EVT) to recognize the unknown samples misclassified into known classes. Specifically, in view of the existence of the domain gaps, we propose a Distance-Rectified Weibull model based on EVT, which can effectively reduce the open set risk, thus keep balance in OSDA. As illustrated in Fig. 1, the domain gaps are minimized by centroid alignment (see Fig. 1 (b)). Unfortunately, an unknown sample (gray arrow) is misclassified into a known class by the classifier. As the centroid deviation angle of this sample (θ_3) is extremely larger that it exceeds a threshold ζ, this sample is statistically rejected as the unknown sample (see Fig. 1 (c)).

In addition, to learn better centroids, we propose an improved centroid update strategy based on sample reweighting and adaptive update rate. This strategy can cooperate well with the centroid alignment.

Our contributions are summarized as follows:

(1) We propose a balanced OSDA method via centroid alignment. The features of both domains are first encoded into a hypersphere. Then, we bound the centroid deviation angles to enhance the domain-invariance and discrimination of the representations.

(2) We propose a Distance-Rectified Weibull model based on EVT to recognize the unknown samples according to the centroid deviation angles of the target features. Experimental results show that this strategy can significantly improve the performance of unknown class samples being correctly recognized. As a consequence, our method could significantly reduce the open set risk.

(3) We propose a new centroid update strategy to cooperate with centroid alignment based on sample reweighting and adaptive update rate.

Related Work

Open Set Domain Adaptation. In OSDA, since the target unknown samples can easily be confused with the target known samples, they will mislead the alignment between domains. Some methods try to use a binary classifier to filter the unknown samples from the target domain, and then only use the target known samples to align with the source domain (Panareda Busto and Gall 2017; Shermin et al. 2020; Feng et al. 2019; Silvia Bucci 2020; Pan et al. 2020). For example, Liu et al. (Liu et al. 2019) train a multi-binary classifier to progressively separate the samples of unknown and known classes. You et al. (You et al. 2019) quantify the sample-level transferability and recognize the unknown samples based on the transferability. Different from these methods, our method uses EVT to recognize the unknown samples according to the centroid deviation angles. Another line of works, e.g., OSBP (Saito et al. 2018b), use the adversarial learning to increase the prediction variances so that the generator can choose to align a target sample with the source known or reject it as an unknown target samples.

Extreme Value Theory. As a branch of statistics, Extreme Value Theory (EVT) (Kotz and Nadarajah 2000) is used to analyze and model the distribution of abnormally low or abnormally high values of data. For example, Bendale et al. (Bendale and Boult 2016) propose a EVT based meta-recognition method OpenMax to address open set recognition on the basis of softmax predictions of a classifier. Scheirer et al. (Scheirer, Jain, and Boult 2014) propose a Weibull-calibrated SVM (W-SVM) to combine EVT for score calibration with two separated SVMs.

The Proposed Method

Problem Definition

The OSDA problem involves two domains: the labeled source domain $\mathcal{S} = \{x_s, y_s | x_s \in X_s, y_s \in Y_s\}$ and the unlabeled target domain $\mathcal{T} = \{x_t | x_t \in X_t\}$, where the two domains share C classes, i.e., the known classes. In addition, the target domain contains classes that are not available in the source domain. We regard these samples as the class $C+1$, i.e., the unknown class. Samples in the source domain $\mathcal{S}$ and the target domain $\mathcal{T}$ are drawn from different probabilities $p(x_s)$ and $p(x_t)$, respectively. In both open set and closed set settings, $p(x_s) \neq p(x_t)$. The goal of our method is to learn the domain-invariant latent representations z_s, z_t and an adaptive classifier for recognizing the C known classes in target domain and rejecting the unknown samples simultaneously. The main idea is illustrated in Fig. 2.

Hyperspherical Variational Auto-Encoders

In order to obtain the domain-invariant hyperspherical representations, we employ $\mathcal{S}$-VAE (Davidson et al. 2018) to encode the data samples into a hyperspherical latent space. Different from the vanilla VAE, $\mathcal{S}$-VAE uses the von Mises-Fisher (vMF) distribution as the prior and posterior distributions. Using the vMF prior instead of the Gaussian prior can avoid the limitations of *origin gravity* and *soap bubble effect* (Davidson et al. 2018) so as to improve the discrimination of the representations. The vMF distribution is

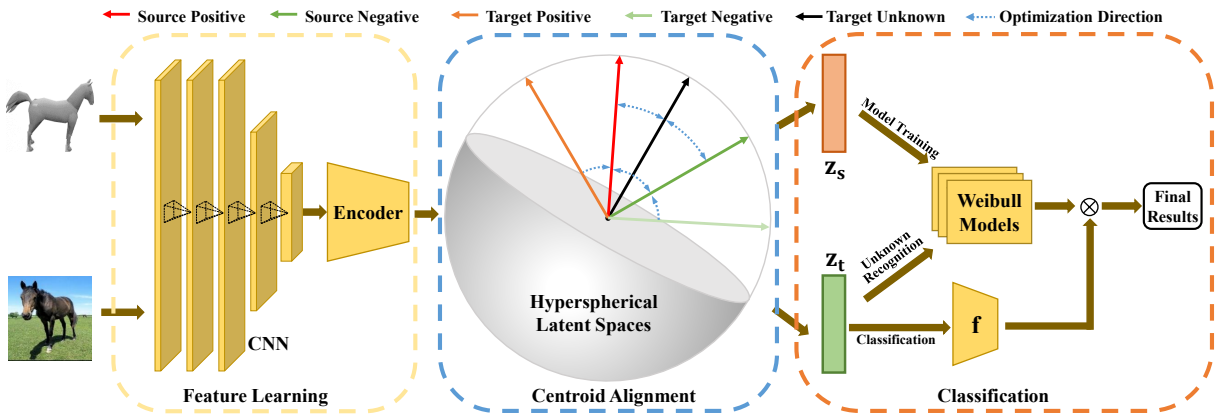


Figure 2: Method overview. The dashed arrows represent the optimization directions, e.g., the arrow between (red,black) means we optimize features to maximize distances between target unknown and source positive. Our method encodes deep features into a hyperspherical latent space. In this space, we bound the centroid deviation angles so that the inter-domain divergence and the intra-domain compactness are maximized simultaneously. With the bounded centroid deviation angles, we train a Distance-Rectified Weibull model for each known class to recognize the misclassified unknown samples.

equivalent to the Gaussian distribution defined on the $(d-1)$-dimensional hypersphere in $\mathbb{R}^d$. Its probability density function w.r.t a d-dimensional random vector z is:

$$q_\theta(z|\mu,\kappa) = \mathcal{C}_d(\kappa)\exp(\kappa\mu^T z) \tag{1}$$

$$\mathcal{C}_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2}\mathcal{I}_{d/2-1}(\kappa)}, \tag{2}$$

where $\mu \in \mathbb{R}^d$ represents the mean direction and $\kappa \in \mathbb{R}$ represents the concentration degree around μ, $||\mu|| = 1$, $\kappa > 0$. $\mathcal{C}_d(\kappa)$ is the normalizing constant, and $\mathcal{I}_u$ denotes the modified Bessel function of the first kind at order u. To obtain the latent representations, we use an encoder to output μ and κ, and then adopt the reparametrization sampling schemes to obtain z, i.e., $q(z|x) = q(z|\mu(x), \kappa(x))$. We train $\mathcal{S}$-VAE by minimizing the following loss:

$$\mathcal{L}_{vae} = \mathcal{L}^s_{vae} + \mathcal{L}^t_{vae}, \tag{3}$$

where

$$\mathcal{L}^s_{vae} = -\,\mathrm{KL}[q_\theta(z_s|x_s)||q(z_t|x_t)] + \mathbb{E}_{q_\theta(z_s|x_s)}[\log p_\phi(x_s|z_s)],$$
$$\mathcal{L}^t_{vae} = -\,\mathrm{KL}[q_\theta(z_t|x_t)||q(z_s|x_s)] + \mathbb{E}_{q_\theta(z_t|x_t)}[\log p_\phi(x_t|z_t)],$$

where $\mathrm{KL}(\cdot)$ is the Kullback-Leibler divergence, $q(z|x)$ and $p(x|z)$ are approximated by an encoder parameterized by θ and a decoder parameterized by ϕ, respectively. Notably, we take the source representations z_s and the target representation z_t as the prior of each other to enhance the domain-invariance of the representations.

With the source representations and labels, we train a classifier f by optimizing the cross entropy loss:

$$\mathcal{L}_{cls} = \frac{1}{N_s}\sum_{i=1}^{N_s} \ell_{ce}(f(z_s^i), y_s^i) \tag{4}$$

Centroid Alignment on Hypersphere

Generally, the centroid of a class can well represent the direction of the whole class. Therefore, we could reduce the domain gaps by aligning the centroids of the two domains. However, in the setting of OSDA, the target domain is unlabeled. The pseudo-labels of the target domain are also unreliable and may mislead the alignment of the two domains. Therefore, we use the centroids of the source domains as the centroids of both domains, and then enforce all samples progressively approach their centroids. We define the angular distance between a sample and a centroid as centroid deviation angle. Then, we employ centroid alignment to bound the centroid deviation angles. We try to achieve the following three goals:

Goal1: *In the source domain, the intra-class representations cluster more tightly while the inter-class representations distribute more dispersedly.*

We achieve **Goal1** by minimizing the following loss:

$$\mathcal{L}_{ca1} = \frac{1}{N_s}\sum_{i=1}^{N_s} \max[m + d(z_s^i, c_s^{y_s^i}) - \min_{j\neq y_i} d(z_s^i, c_s^j), 0], \tag{5}$$

where $d(z_1, z_2) = \arccos(z_1 \cdot z_2)$ is a function to compute the angular distance, N_s is the number of source samples, c_s^v denotes the source centroid of class v, m is the angular margin between two representations.

Achieving **Goal1** can increase the intra-class density and maximize the inter-class boundaries of the source domain.

Goal2: *The target domain representations could align with the source domain representations at both the class-level and sample-level.*

We deploy the following loss to achieve **Goal2**:

$$\mathcal{L}_{ca2} = \frac{1}{C}\sum_{i=1}^{C} \max[m + d(c_t^i, c_s^i) - \min_{j\neq i} d(c_t^i, c_s^j), 0] + \frac{1}{N_t^C}\sum_{i=1}^{N_t^C} \max[m + d(z_t^i, c_s^{y_t^i}) - \min_{j\neq y_t^i} d(z_t^i, c_s^j), 0] \tag{6}$$

where the first and second term are the class-level alignment loss and sample-level alignment loss, respectively. N_t^C is the

number of known class samples. c_t^v denotes the target centroid of class v.

Note that as the target domain is unlabeled, we use the classifier f to get pseudo-labels for the target samples.

Goal3: *The unknown class representations go farther away from centroids of all the known classes.*

We achieve **Goal3** through the following margin loss:

$$\mathcal{L}_{unk} = \frac{1}{N_t^{C+1}C} \sum_{i=1}^{N_t^{C+1}} \sum_{j=1}^{C} \max[m - d(z_t^{u(i)}, c_s^j), 0], \quad (7)$$

where $z_t^{u(i)}$ denote the i-th unknown sample and N_t^{C+1} denotes the number of the unknown samples.

Since the unknown samples would hinder the process of domain alignment and even lead to negative transfer (Pan and Yang 2010), achieving **Goal3** could circumvent this negative effect.

Achieving the three goals is of double significance. On one hand, we can learn the domain-invariant and discriminative latent representations. On the other hand, the centroid deviation angles of the target representations are bounded, making it feasible to recognize the target unknown samples based on EVT. The t-SNE in Fig. 3 verify that our method can effectively achieve the three goals.

Then, the objective for aligning the two domains on the hypersphere is formulated as follows:

$$\mathcal{L}_{align} = \mathcal{L}_{ca1} + \mathcal{L}_{ca2} + \mathcal{L}_{unk} \quad (8)$$

Finally, the overall objective is defined as:

$$\mathcal{L} = \mathcal{L}_{vae} + \lambda \mathcal{L}_{align} + \gamma \mathcal{L}_{cls} \quad (9)$$

Unknown Samples Recognition

Since there are only known samples in the source domain, the classifier f trained on the source domain tends to mistakenly classify a target unknown feature into one of the known classes though its classification space includes the unknown class. In the hyperspherical latent space, we assume that if a target unknown sample z_t^u is misclassified into class k, then the centroid deviation angle of it with the source centroid of class k may be distinctly larger than that of the true sample of class k with their centroid, i.e., $d(z_t^k, c_s^k) < d(z_t^u, c_s^k)$. Motivated by (Bendale and Boult 2016; Scheirer, Jain, and Boult 2014), we propose to adopt the Weibull distribution to model the distribution of the centroid deviation angles. Then, any representation with abnormally large centroid deviation angles will be predicted to be unknown class.

However, although there are many EVT-based open set recognition methods (Bendale and Boult 2016; Scheirer, Jain, and Boult 2014), these methods cannot be directly used in the OSDA problem. Due to the existence of the domain gaps, the centroid deviation angles of a target known sample z_t^k with the source centroid of class k is larger than that of the source sample of class k, i.e., $d(z_s^k, c_s^k) < d(z_t^k, c_s^k)$. Therefore, using the existing EVT-based open set recognition methods, e.g., OpenMax (Bendale and Boult 2016) and WSVM (Scheirer, Jain, and Boult 2014), in OSDA will lead to a large number of target known samples being misclassified as unknown. This is proved by the experimental results in Ablation Study.

Algorithm 1 Unknown Samples Recognition Using EVT

Training Phase:

Input: The source representations $\{z_s\}$ and their centroids: $\{c_s\}$, tail size η, classifier f.

1: Classify $\{z_s\}$ and get the correctly classified ones $\{\tilde{z}_s\}$.

for k = 1 to C do

2: $\forall\, \tilde{z}_s^k \in$ class k, compute its centroid deviation angle $d_k = \arccos(\tilde{z}_s^k \cdot c_s^k)$ and get the distance set of class k $D_k = \{d_k\}$

3: Fit Weibull model $\mathcal{M}_k$ with D_k and tail size η.

end for

Output: Weibull model set $\{\mathcal{M}\}$ for all known classes.

Recognizing Phase:

Input: The target representation z_t, Weibull model set $\{\mathcal{M}\}$, classifier F, the set of source centroid $\{c_s\}$, the set of target centroid $\{c_t\}$, threshold ζ.

1: Classify z_t to get its prediction $\hat{y}_t = k$.

2: Compute the rectified distance $\tilde{d}_k$ according to Eq. (11).

3: Compute the probability of z_t belonging to unknown class: $\omega_k = \mathcal{M}_k(\tilde{d}_k)$ according to Eq. (10).

4: Modify the prediction $\hat{y}_t$ = C+1 if $\omega_k > \zeta$.

In view of the domain gaps, we propose a Distance-Rectified Weibull model for OSDA. Specifically, for a target sample z_t pseudo-labeled as class k, the probability of z_t belonging to the unknown class can be expressed as:

$$\omega_k = 1 - \exp\left(-\left(\frac{\tilde{d}(z_t, c_s^k) - \nu_k}{\sigma_k}\right)^{\tau_k}\right) \quad (10)$$

where ν_k, σ_k and τ_k are three parameters for the Weibull model of class k. $\tilde{d}(z_t, c_s^k)$ is the rectified angular distance which can be computed as:

$$\tilde{d}(z_t, c_s^k) = \max[d(z_t, c_s^k) - d(c_t^k, c_s^k), 0] \quad (11)$$

We train multiple Weibull models, one for each class. Then, for a target representation z_t predicted to be class k, we use the Weibull model of class k to estimate its unknown probability ω_k. If ω_k is larger than a prior threshold ζ, the prediction of z_t will be modified to $C+1$.

For a clear understanding, we present the process of the unknown samples recognition in **Algorithm 1**.

Noteworthily, in the model training phase, we only use the correctly classified source representations to train the Weibull models, this is in line with (Bendale and Boult 2016). Moreover, the tail size η controls the ratio of the extreme value in the distribution, we report the sensitivity of our method w.r.t. η in Parameter Sensitivity Analysis.

Adaptive Centroid Update Strategy

Xie et al. (Xie et al. 2018) propose a centroid update strategy, the centroid of class k in iteration n is updated as follows:

$$c_l^k(n) = \frac{1}{N^k} \sum_{i=1}^{N^k} x_i^k, c_g^k(n) = (1-\alpha) c_g^k(n-1) + \alpha c_l^k(n)$$

where $c_l^k(n)$ and $c_g^k(n)$ are the local and global centroid of class k, respectively. N^k is the number of samples belonging to class k in iteration n, α is the update rate.

However, there are two limitations in Xie's strategy. First, it does not consider the number of samples of different classes in a mini batch. All classes, regardless of their quantity, are updated at a same rate α. For classes with few or even only one sample, the risk of being misled by the misclassified samples will increase. Second, it does not consider the confidence score of the classifier for each target sample. If a target sample is classified into a class with very low confidence, this sample is quite likely to be wrongly classified.

To alleviate these limitations, we improve Xie's strategy by reweighting samples and adopting adaptive update rate. Specifically, we use entropy to quantify the prediction of the classifier: $H(p) = -p\log p$, where p is the confidence score of a sample predicted to a class, and compute a weight w for each target sample: $w = 1 + e^{-H(p)}$. Then the local centroid of class k in iteration n can be updated as $c_l^k(n) = \frac{1}{S^k}\sum_{i=1}^{N^k} w_i^k z^{k(i)}$, where w_i^k is the weight for i-th sample of class k, $S^k = \sum_{i=1}^{N^k} w_i^k$ is the sum of weights for all samples predicted to be class k, $z^{k(i)}$ is the i-th sample of class k.

In addition, we compute a scale factor r to adaptively adjust the update rate for each class which considers the sample quantity of a class: $r = \frac{S^k C}{S}$, where S is the sum of weights for all samples in a mini-batch. Therefore, the more samples belonging to a class or the higher their confidence scores, the larger their scale factors are. Finally, the global centroid is updated as:

$$c_g^k(n) = (1 - r\alpha)c_g^k(n-1) + r\alpha c_l^k(n). \tag{12}$$

Note that the source samples have the ground-truth labels, so we fix $w = 1$ for all the source samples.

Theoretical Insight

Theorem 1. Open Set Domain Adaptation Theory (Luo et al. 2020). Given a hypothesis $\mathcal{H}$ with a mild condition that constant function $C+1 \in \mathcal{H}$, then for any $h \in \mathcal{H}$, the expected error on the target domain can be bounded by:

$$\begin{aligned}\frac{R_t(h)}{1-\pi_{C+1}^t} \leq & R_s(h) + D_{\mathcal{H}}(p(z_s), p(z_t)) + \lambda \\ & + \underbrace{\frac{\pi_{C+1}^t}{1-\pi_{C+1}^t} R_{t,C+1}}_{\text{open set risk}},\end{aligned} \tag{13}$$

where $\lambda = \min_{h \in H} R_s(h) + R_t^*(h)$ is the adaptability of the known samples between domains (Ben-David et al. 2010), $R_s(h)$ and $R_t(h)$ are the expected error on the source and target domain, respectively. $D_{\mathcal{H}}(p(z_s), p(z_t))$ denotes the discrepancies between two distributions. $\pi_{C+1}^t = p(z_t = C+1)$ is the class-prior probability. $R_{t,C+1}(h)$ is the risk of h on the unknown class.

In Theorem 1, the source expected error $R_s(h)$ is minimized by $\mathcal{L}_{cls}$, the domain discrepancies $d_H(p(z_s), p(z_t))$ are reduced by $\mathcal{L}_{align}$. In addition, as we employ the EVT-based Weibull model to enhance the ability of recognizing the unknown samples, $R_{t,C+1}(h)$ is minimized. Correspondingly, the open set risk is reduced. As for the domain adaptability λ, usually, it is considered sufcient low (Saito et al. 2018a; Long et al. 2018a). Otherwise, one should consider choosing a more related source domain for adaptation. Therefore, the proposed method could bound the expected error on the target domain theoretically.

Experiment

Evaluation Protocol

In line with the previous work (Silvia Bucci 2020), we evaluate all the compared methods with three metrics, i.e., the average accuracy for known classes **OS***, the accuracy for the unknown class **Unk** and the harmonic mean accuracy $\mathbf{H} = 2 \times \text{OS}^* \times \text{Unk}/(\text{OS}^* + \text{Unk})$. **H** is a balanced evaluation metric which correctly assesses the performance of the methods on both known and unknown class samples.

Datasets and Compared Methods

Office-31 (Saenko et al. 2010) includes 31 classes from 3 domains: **A**, **W** and **D**. Following (Saito et al. 2018b), we select 10 classes as known and 11 classes as unknown.
VisDA-2017 (Peng et al. 2017) contains 2 domains: **Synthetic** and **Real**. Each domain includes 12 classes. Following (Saito et al. 2018b), we take the first 6 classes as known and the remaining as unknown.
Image-CLEF[1] includes 4 domains: **I**, **C**, **P** and **B**. Each domain contains 12 classes. We use the first 6 classes in alphabetical order as the known and the rest as the unknown.

The compared methods include: **ROS**(Silvia Bucci 2020), **OSBP** (Saito et al. 2018b), **STA** (Liu et al. 2019), **UAN** (You et al. 2019), **DAOD** (Fang et al. 2020) and **CDAN** (Long et al. 2018b). All the methods are proposed in recent 3 years.

Implementation Details

For the model structrue, the encoder of the $\mathcal{S}$-VAE includes 3 fully-connected (FC) layers, its output is activated by softplus. The decoder is composed of 2 FC layers. The output of its first FC layer is activated by ReLU and then fed into the second FC layer. The classifier contains only one FC layer, and its output is processed by Logsoftmax. We adopt adam (JLB 2015) to optimize these models with learning rate 4e-4 for $\mathcal{S}$-VAE models and 1e-3 for the classifier. All the learning rate decreases during the training following an inverse decay scheduling.

As for the hyperparameters, we get the optimal hyperparameters through importance-weighted cross-validation (Sugiyama, Krauledat, and MÃžller 2007). As our method performs stably under some hyperparameters, we fix the centroid update rate $\alpha = 0.2$, the tail size $\eta = 0.02$, the threshold $\zeta = 0.98$, and the margin angle $m = 90°$ across all the experiments. In addition, for Office-31 and Image-CLEF, we set $\lambda = 1.0$, $\gamma = 1.0$. For VisDA-2017, we set $\lambda = 0.5$, $\gamma = 0.5$.

[1] http://imageclef.org/2014/adaptation

Method	A→W			A→D			W→A			W→D			D→A			D→W			Avg.		
	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H
UAN	**95.5**	31.0	46.8	**95.6**	24.2	38.9	**94.1**	38.8	54.9	81.5	41.4	53.0	93.5	53.4	68.0	**99.8**	93.0	96.0	93.4	40.3	55.1
STA	92.1	58.0	71.0	95.4	45.5	61.6	92.1	46.2	60.9	96.6	48.5	64.4	**94.1**	55.0	69.4	97.1	49.7	65.5	**94.6**	50.5	65.5
OSBP	86.8	79.2	82.7	90.5	75.5	82.4	73.0	74.4	73.7	99.1	84.2	91.1	76.1	72.3	75.1	97.7	96.7	**97.2**	87.2	80.4	83.7
ROS	88.4	76.7	82.1	87.5	77.8	82.4	69.7	86.6	77.2	**100.0**	99.4	**99.7**	74.8	81.2	77.9	99.3	93.0	96.0	86.6	85.8	85.9
Ours	88.8	**92.6**	**90.7**	92.4	**88.6**	**90.5**	81.9	**98.3**	**89.4**	96.5	**100.0**	98.2	82.9	**92.0**	**87.2**	89.3	**98.5**	93.7	88.6	**95.0**	**91.6**

Table 1: Accuracy (%) on Office-31 with ResNet-50 as backbone. The best results are highlighted by bold numbers.

Method	B→C			B→I			B→P			C→B			C→I			C→P		
	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H
DAOD	79.4	82.0	80.7	78.4	**90.9**	84.2	72.1	**80.8**	**76.2**	51.3	47.1	49.1	79.0	88.6	83.6	74.5	**78.9**	76.6
STA	93.3	51.7	66.5	86.0	60.7	71.2	77.7	48.7	59.8	61.3	69.7	65.2	91.7	66.7	77.2	84.0	54.0	65.7
OSBP	87.0	81.0	83.9	85.3	65.7	74.2	66.3	66.7	66.5	**62.0**	58.0	59.9	89.0	80.0	84.3	**87.7**	53.7	66.6
ROS	78.3	90.0	83.8	73.0	76.3	74.6	59.0	67.3	62.9	59.0	68.3	63.3	78.3	83.0	80.6	68.7	78.7	73.3
Ours	**95.7**	**98.3**	**97.0**	**87.7**	86.7	**87.2**	**80.3**	66.0	72.5	57.0	**86.7**	**68.8**	**94.7**	**92.3**	**93.5**	76.7	76.7	**76.7**

	I→B			I→C			I→P			P→B			P→C			P→I			Avg.		
	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H	OS*	Unk	H
DAOD	54.5	56.9	55.7	80.3	82.0	81.2	73.3	80.8	76.9	51.7	51.0	51.3	79.0	82.0	80.5	79.6	88.6	83.9	71.1	75.8	73.3
STA	**62.3**	54.0	57.9	94.0	53.7	68.4	80.7	59.0	68.2	**61.3**	43.7	51.0	93.7	47.7	63.2	**90.0**	51.0	65.1	**81.3**	55.1	65.0
OSBP	55.7	60.7	58.1	80.7	92.7	86.3	66.3	74.3	70.1	52.3	61.0	56.3	94.0	68.0	78.9	66.0	80.7	72.6	74.4	70.2	71.5
ROS	58.0	59.7	58.8	88.7	92.7	90.6	78.0	76.0	77.0	47.3	59.3	52.7	71.3	90.3	79.7	79.7	81.3	80.5	69.9	76.9	73.1
Ours	54.0	**82.7**	**65.3**	**95.7**	**97.3**	**96.5**	**81.3**	**88.7**	**84.8**	55.7	**70.3**	**62.1**	**94.7**	**91.3**	**93.0**	84.0	**91.7**	**87.7**	79.8	**85.7**	**82.1**

Table 2: Accuracy (%) on Image-CLEF with ResNet-50 as backbone. The best results are highlighted by bold numbers.

Method	bcyc	bus	car	mcyc	train	truck	OS*	Unk	H
CDAN	49.6	58.1	**71.4**	86.6	**85.6**	12.3	60.6	0	0
STA	38.2	69.1	51.2	87.6	78.0	11.1	55.9	75.2	64.1
OSBP	53.9	**77.6**	56.4	**89.1**	74.4	22.2	62.3	71.3	66.5
Ours	**77.7**	72.0	48.8	82.2	81.5	**34.7**	**66.2**	**94.1**	**77.7**

Table 3: Accuracy (%) on VisDA-2017 with ResNet-50 as backbone. The best results are highlighted by bold numbers.

Settings	A→W					
	OS^*_{acc}	Unk_{acc}	H_{acc}	OS^*_{prc}	Unk_{prc}	H_{prc}
w/o CA, UnkR	90.5	50.6	64.9	68.9	88.3	77.4
w/o UnkR	92.6	84.8	88.5	87.3	91.9	89.6
w/o CA	90.8	67.7	77.6	78.8	87.5	82.9
w/o DR	82.2	98.9	89.8	98.1	82.1	89.4
Ours	88.8	92.6	90.7	93.4	87.4	90.3

Table 4: Ablation study: accuracy (%) and precision (%) of the proposed method and its four variants. CA, UnkR and DR denote centroid alignment, unknown recognition and distance rectification, respectively.

For all the compared methods, we either report the results of the original papers if the results are tested under the same setting, or the best results we can achieve.

Experimental Results

We report the experimental results on three datasets in Table 1-3. The backbone network is ResNet-50.

On Office-31, we observe that the harmonic mean accuracies of our method outperform all the compared methods on 5 out of 6 tasks. The only exception is W→D, on which the proposed method still achieves the second best results. Notebly, UAN and STA achieve the best OS* on 5 out of 6 tasks. However, due to their low classification performance for the unknown samples, their harmonic mean accuracies are very low. Obviously, the classification abilities of UAN and STA are biased. As a comparison, our method is no doubt more balanced: the average accuracy for the unknown samples is 95.0%, which is 9.2% higher than the second best method ROS. Simultaneously, our method still maintains very high classification performance for the known samples, which makes the average harmonic mean accuracies of our method exceed at least 5.7% of the compared methods. Similar trends could be observed on Image-CLEF and VisDA.

In Table 3, we report results of a closed set method, i.e., CDAN. We observe that CDAN misclassify all the unknown samples into known ones, which reveals that the closed set method would fail in the open set setting.

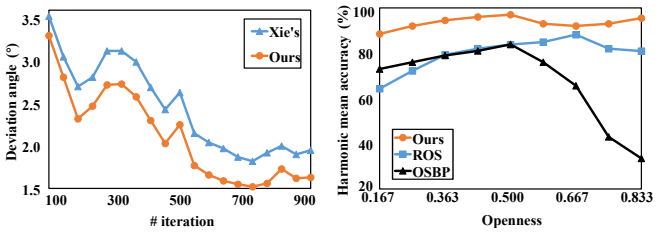
Ablation Study

We conduct ablation study on A→W to evaluate contributions of different components in our method and report results in Table 4. Apart from the results based on the accuracy metric, we also report the results based on the precision metric, which is computed as: $\text{precision} = \frac{TP}{TP+FP}$, i.e., of the samples predicted to be class k, the proportion of samples that actually belong to class k.

From Table 4, we can make the following observations: (1) With Centroid Alignment (CA), OS^*_{acc} and Unk_{acc} are improved compared with the setting without CA, which verifies that CA could not only reduce the domain gaps but also increase the separation between known and unknown. (2) In the setting without CA, the accuracy for unknown samples

(a) ResNet-50 (b) OSBP (c) ROS (d) Ours

Figure 3: The t-SNE visualization of the source and target representations on D→A.



(a) (b)

Figure 4: (a) Comparison of the deviation angles between the learned centroids and the ground-truth centroids under different update strategies. (b) Experimental results w.r.t. varying openness on B→C.

is only 67.7%, while it increases to 92.6% after adding CA. These results indicate that UnkR requires CA as a prerequisite. UnkR alone does not have a strong unknown samples recognition ability. Without CA to enlarge the angular distances of known and unknown, the bound of known and unknown will not be so clear, making it difficult for UnkR to recognize unknown samples correctly. (3) Without distance rectification, OS^*_{acc} and Unk_{prc} decrease, while Unk_{acc} and OS^*_{prc} increase, which means that many known samples are misclassified into unknown samples. This reveals that Weibull models trained with source samples cannot be directly applied to the target samples due to the existence of domain gaps.

Therefore, all the components in the proposed method are effective and indispensable.

Feature Visualization

We plot t-SNE of D→A in Fig. 3 to visualize the distribution of different representations. From Fig. 3, we observe that the representations learned by our method are closely clustered together (**Goal1**). Secondly, the representations belonging to the same class from the source domain (in red) and the target domain (in blue) are closely distributed together, while the features of different classes have very clear boundaries (**Goal2**). Thirdly, the target unknown representations (in gray) distribute far away from the known representations (**Goal3**). Therefore, of all the methods, only our method successfully achieves **Goal1**-**Goal3** simultaneously.

Effectiveness of the New Centroid Update Strategy

We plot the curves of the average deviation angles between the centroids learned by different strategies (Xie's and ours) and the ground-truth centroids across all the known classes in Fig. 4 (a). With the increase of the iterations, the centroids learned by our method always have smaller deviation angles than the centroids learned by Xie's strategy, which verifies that our method could align two domains more accurately.

Robustness to Different Openness

To evaluate the robustness of our method with different openness, we report results of three methods with openness varying from 0.167 to 0.833 in Fig. 4 (b). The openness is the ratio of unknown classes to all classes in the dataset. We observe that our method can keep high performance under different levels of openness. OSBP suffers from performance degradation when the openness approaches 1. ROS performs better than OSBP under large openness, but its performance will reduce when the openness is nearly 0.

Limited by space, we present more experimental analyses in **Supplementary Material**.

Conclusion

In this paper, we propose a balanced OSDA method based on centroid alignment in the hyperspherical latent space. We propose to bound the centroid deviation angles to ensure that the learned representations are not only domain invariant, but also discriminative. With the bounded centroid deviation angles, we further propose a Distance-Rectified Weibull model based on EVT to recognize the unknown samples misclassified into known classes, which can reduce the open set risk. In addition, we propose an improved centroid update strategy to cooperate with centroid alignment. The experimental results verify the effectiveness of our method.

Acknowledgments

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

References

Baktashmotlagh, M.; Faraki, M.; Drummond, T.; and Salzmann, M. 2018. Learning Factorized Representations for Open-set Domain Adaptation. In *ICLR*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine learning* 79(1-2): 151–175.

Bendale, A.; and Boult, T. E. 2016. Towards Open Set Deep Networks. In *CVPR*.

Davidson, T. R.; Falorsi, L.; De Cao, N.; Kipf, T.; and Tomczak, J. M. 2018. Hyperspherical variational auto-encoders. In *UAI*, 856–865. Association For Uncertainty in Artificial Intelligence (AUAI).

Fang, Z.; Lu, J.; Liu, F.; Xuan, J.; and Zhang, G. 2020. Open set domain adaptation: Theoretical bound and algorithm. *TNNLS* .

Feng, Q.; Kang, G.; Fan, H.; and Yang, Y. 2019. Attract or distract: Exploit the margin of open set. In *ICCV*, 7990–7999.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 1989–1998. PMLR.

Jing, M.; Zhao, J.; Li, J.; Zhu, L.; Yang, Y.; and Shen, H. T. 2020. Adaptive Component Embedding for Domain Adaptation. *IEEE Transactions on Cybernetics* .

JLB, D. P. K. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Kotz, S.; and Nadarajah, S. 2000. *Extreme value distributions: theory and applications*. World Scientific.

Li, J.; Chen, E.; Ding, Z.; Zhu, L.; Lu, K.; and Shen, H. T. 2020. Maximum Density Divergence for Domain Adaptation. *TPAMI* .

Li, J.; Jing, M.; Lu, K.; Zhu, L.; and Shen, H. T. 2019. Locality preserving joint transfer for domain adaptation. *TIP* 28(12): 6103–6115.

Li, J.; Lu, K.; Huang, Z.; Zhu, L.; and Shen, H. T. 2018a. Heterogeneous domain adaptation through progressive alignment. *TNNLS* 30(5): 1381–1391.

Li, J.; Lu, K.; Huang, Z.; Zhu, L.; and Shen, H. T. 2018b. Transfer independently together: A generalized framework for domain adaptation. *IEEE transactions on cybernetics* 49(6): 2144–2155.

Liu, H.; Cao, Z.; Long, M.; Wang, J.; and Yang, Q. 2019. Separate to adapt: Open set domain adaptation via progressive separation. In *CVPR*, 2927–2936.

Long, M.; Cao, Y.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018a. Transferable representation learning with deep adaptation networks. *TPAMI* 41(12): 3071–3085.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018b. Conditional adversarial domain adaptation. In *NeurIPS*, 1640–1650.

Luo, Y.; Wang, Z.; Huang, Z.; and Baktashmotlagh, M. 2020. Progressive Graph Learning for Open-Set Domain Adaptation. In *ICML*.

Pan, S. J.; and Yang, Q. 2010. A survey on transfer learning. *TKDE* 22(10): 1345–1359.

Pan, Y.; Yao, T.; Li, Y.; Ngo, C.-W.; and Mei, T. 2020. Exploring Category-Agnostic Clusters for Open-Set Domain Adaptation. In *CVPR*, 13867–13875.

Panareda Busto, P.; and Gall, J. 2017. Open set domain adaptation. In *ICCV*, 754–763.

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

Prettenhofer, P.; and Stein, B. 2010. Cross-language text classification using structural correspondence learning. In *ACL*, 1118–1127.

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

Saito, K.; Watanabe, K.; Ushiku, Y.; and Harada, T. 2018a. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 3723–3732.

Saito, K.; Yamamoto, S.; Ushiku, Y.; and Harada, T. 2018b. Open set domain adaptation by backpropagation. In *ECCV*, 153–168.

Scheirer, W. J.; Jain, L. P.; and Boult, T. E. 2014. Probability models for open set recognition. *TPAMI* 36(11): 2317–2324.

Shermin, T.; Lu, G.; Teng, S. W.; Murshed, M.; and Sohel, F. 2020. Adversarial Network with Multiple Classifiers for Open Set Domain Adaptation. *TMM* .

Silvia Bucci, Mohammad Reza Loghmani, T. T. 2020. On the Effectiveness of Image Rotation for Open Set Domain Adaptation. In *ECCV*.

Sugiyama, M.; Krauledat, M.; and MÃžller, K.-R. 2007. Covariate shift adaptation by importance weighted cross validation. *JMLR* 8(May): 985–1005.

Xie, S.; Zheng, Z.; Chen, L.; and Chen, C. 2018. Learning semantic representations for unsupervised domain adaptation. In *ICML*, 5423–5432.

You, K.; Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2019. Universal domain adaptation. In *CVPR*, 2720–2729.