Interpretable Open-Set Domain Adaptation via Angular Margin Separation

Xinhao Li[1], Jingjing Li[1,2(✉)], Zhekai Du[1], Lei Zhu[3], and Wen Li[1]

[1] University of Electronic Science and Technology of China
Mc1114207918@outlook.com, lijin117@yeah.net, dzk1996411@163.com,
liwenbnu@gmail.com
[2] Institute of Electronic and Information Engineering of UESTC in Guangdong
[3] Shandong Normal University
leizhu0608@gmail.com

Abstract. Open-set Domain Adaptation (OSDA) aims to recognize classes in the target domain that are seen in the source domain while rejecting other unseen target-exclusive classes into an unknown class, which ignores the diversity of the latter and is therefore incapable of their interpretation. The recently-proposed Semantic Recovery OSDA (SR-OSDA) brings in semantic attributes and attacks the challenge via partial alignment and visual-semantic projection, marking the first step towards interpretable OSDA. Following that line, in this work, we propose a representation learning framework termed Angular Margin Separation (AMS) that unveils the power of discriminative and robust representation for both open-set domain adaptation and cross-domain semantic recovery. Our core idea is to exploit an additive angular margin with regularization for both robust feature fine-tuning and discriminative joint feature alignment, which turns out advantageous to learning an accurate and less biased visual-semantic projection. Further, we propose a post-training re-projection that boosts the performance of seen classes interpretation without deterioration on unseen classes. Verified by extensive experiments, AMS achieves a notable improvement over the existing SR-OSDA baseline, with an average 7.6% increment in semantic recovery accuracy of unseen classes in multiple transfer tasks. Our code is available at AMS.

Keywords: open-set domain adaptation · zero-shot learning

1 Introduction

The advent of deep neural network (DNN) [15] and corresponding deep learning algorithms [17] has enabled computer vision unprecedented development and wide application in real-world production. However, meanwhile, some common assumptions adopted by conventional machine learning frameworks, such as the i.i.d assumption and the closed world assumption, have gradually hindered the data-driven large-scale deep learning models, bringing two potential challenges in real-world applications [5,6,49]. First, the distribution of target data faced in deployment may be quite different from that of the well-labeled source data, and it is often too costly and even infeasible to collect adequate annotations for the new data distribution. Second, the real world is an open world, with new classes unseen during training possibly emerging at any time, and failure to carefully handle them could lead to fatal consequences, e.g., for self-driving vehicles.

Under such demands, open-set domain adaptation (OSDA) [31,38,25,3] has been widely studied, bridging the domain gap between a well-labeled source domain and an unlabeled target domain while rejecting all target-exclusive classes into one unknown class. However, few existing methods pay attention to the inter-class difference among different unknown classes, and none of them is capable of their interpretation. Although such deficiencies may not violate the task of OSDA by definition, they are essentially suboptimal for many real-world scenarios, where the unknown classes are diverse and their interpretation could be important for human intervention and model evolution. To overcome this limitation, semantic recovery domain adaptation (SR-OSDA) [13] has been recently proposed, which introduces semantic attributes to interpret the unseen novelties. On top of OSDA, SR-OSDA additionally learns a projection from visual features to their corresponding semantic attributes on seen classes in hope that it could also be applicable to the unseen classes. Besides, SR-OSDA purposefully differentiates between target data detected unknown to avoid interpreting them as one naive unknown class. Nevertheless, SR-OSDA, at its budding phase with only a prospective yet general objective, still has large room to improve.

Extensive studies have shown the significance of representation learning [1,34,21,9] for visual tasks, and we argue that the same is true for SR-OSDA. In this paper, we investigate the power of discriminative and robust representation for both open-set domain adaptation and cross-domain semantic recovery. Specifically, we first exploit an additive angular margin in visual space to fine-tune the pre-trained model on the source domain to learn representation suitable for cross-domain novelty detection and seen class recognition. Afterwards, we resort to additive angular margin again on the visual-semantic joint representation to facilitate compact alignment of seen classes and discriminative separation of unseen classes, which proves effective for both seen class recognition and distinct interpretation of diverse unseen classes. Finally, we propose a post-training re-projection to efficiently boost semantic recovery for seen classes without deterioration on unseen classes. We can summarize our contributions as follows:

- We unveil the power of discriminative and robust representation for SR-OSDA by exploiting an additive angular margin in both the fine-tuning phase and the training phase, which proves advantageous to both open-set domain adaptation and cross-domain semantic recovery.
- We propose a post-training re-projection with minimal cost to further boost semantic recovery on seen classes in the target domain without deterioration on unseen classes.
- Verified by extensive experiments, our proposed AMS achieves a notable improvement over the existing SR-OSDA framework, with an average increment of 7.6% in semantic recovery accuracy of unseen classes, and various improvements on other evaluation metrics in multiple transfer tasks.

2 Related Work

Open-set domain adaptation. Open-set domain adaptation (OSDA) [31,38,25,3,12] is recognized as a more practical form of domain adaptation [7,42,22,18,19]. Confronted with the domain gap between a well-labeled source domain and an unlabeled target domain, OSDA not only needs to recognize the classes in the target domain that are seen in the source domain but also to detect the target-domain-exclusive classes as an unknown (unseen) class. Despite much progress achieved, few works consider the intrinsic diversity of the unknown classes, and none can provide interpretation for them, which in fact could be critical for human intervention or even the evolution of the model. Motivated by such deficiency, semantic-recovery domain adaptation (SR-OSDA) [13] was recently proposed to leverage attributes [16] to recover the semantics of unseen classes and thus realize their interpretation. In this work, we follow the recent SR-OSDA and interpret target unseen classes by semantic attributes.

Zero-shot learning. Zero-shot learning (ZSL) [45,33] aims to enable the machine to recognize classes with no training samples via side information [16,43]. ZSL methods are typically taxonomized by two criteria: (i) inductive ZSL vs. transductive ZSL, and (ii) classical ZSL vs. generalized ZSL (GZSL). Inductive ZSL only has access to the samples and side information of seen classes for training, while transductive ZSL [24,40] can further access the unlabeled samples or side information of the unseen classes. Classical ZSL assumes the sole presence of unseen classes in testing, while GZSL [4,20,44,41] needs to handle both. Since SR-OSDA has access to labeled samples and semantic attributes of seen classes as well as unlabeled samples from unseen classes, and needs to deal with both in inference, it is more similar to transductive GZSL. However, there are two distinct differences worth noting: (i) ZSL does not consider domain gap, and (ii) transductive ZSL typically assumes that the range of unseen classes is known, and thus often uses the semantic attributes of unseen classes with techniques like dictionary learning [48] and matrix factorization [46] to enhance classification. In contrast, SR-OSDA does not assume the range of unseen classes as prior knowledge and aims to recover the semantics of any unseen classes as accurately as possible. Therefore, the semantic attributes of unseen classes are only for evaluation and cannot be used in any form to proactively enhance performance.

3 Method

3.1 Problem Setup

We use $x \in \mathcal{X} \subset \mathbb{R}^p$, $y \in \mathcal{Y}$, and $a \in \mathcal{A} \subset \mathbb{R}^m$ to denote samples, labels, and semantic attributes. Let P, Q be the source and target distribution defined on $\mathcal{X} \times \mathcal{Y}$. In SR-OSDA, the source domain $\mathcal{D}_s$ consists of N_s labeled samples with semantic attribute prototypes $\{x_s^i, y_s^i, a_s^i\}_{i=1}^{N_s}$ with $x_s^i \in \mathcal{X}_s \subset \mathbb{R}^{p \times N_s}$, $y_s^i \in \mathcal{Y}_s$, and $a_s^i \in \mathcal{A}_s \subset \mathcal{R}^{m \times N_s}$ drawn i.i.d from P. The target domain $\mathcal{D}_t$ consists of N_t unlabeled samples $\{x_t^i\}_{i=1}^{N_t}$ with x_t^i drawn i.i.d from Q. Owing to domain gap, $p(x_s) \neq q(x_t)$. The source domain is associated with a set of seen classes

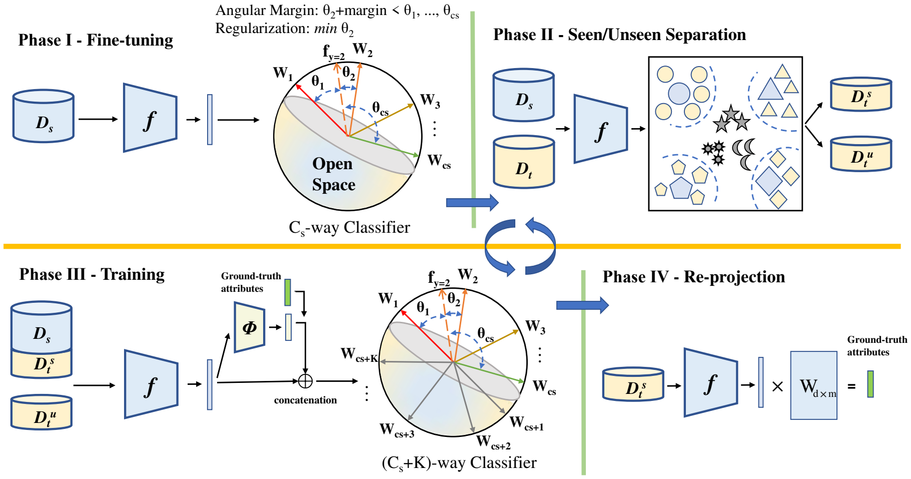


Fig. 1: Illustration of our proposed AMS. Phase I: Fine-tune the pre-trained feature extractor f on the seen classes of source domain with additive angular margin and angular regularization to enable highly discriminative representation of seen classes and leave out open space for future outliers. Phase II: Seen/Unseen separation: Separate target samples into seen and unseen classes based on feature distance and have them pseudo-labeled. Phase III: Train the model on both domains by applying additive angular margin and angular regularization on joint visual-semantic representation. The extra K channels in the classifier work under the maximum-response correspondence (MRC) strategy to differentiate diverse unseen classes. We alternate between Phase II and III to acquire higher accuracy of pseudo-labels. Phase IV: Learn a regularized linear projection W on target samples detected from seen classes to corresponding ground-truth attributes to further boost their semantic recovery. Best viewed in color.

$\mathcal{Y}_s = \{1, ..., |C_s|\}$, which is a subset of classes in target domain $\mathcal{Y}_t = \mathcal{Y}_s \cup \{|C_s| + 1, ...|C_s| + |C_t|\}$. In the m-dimensional semantic space, each class is associated with one semantic attribute prototype. Therefore, $\mathcal{A}_s = \{A_1, ..., A_{|C_s|}\}$ and $\mathcal{A}_t = \mathcal{A}_s \cup \{A_{|C_s|+1}, ..., A_{|C_s|+|C_t|}\}$. Our goal is two-fold: (i) to learn a target prediction function $h_t : x_t \rightarrow y_t$ that correctly classifies the $|C_s|$ known classes and rejects the other classes into an general unseen class, (ii) to learn a visual-semantic projection $\phi_t : x_t \rightarrow a_t$ that recovers the interpretable semantic attributes of target data.

3.2 Framework Overview

When designing a method for SR-OSDA, we face two main challenges: seen/unseen separation, and class differentiation. Primarily, the solving of SR-OSDA largely relies upon the correct separation of seen and unseen classes. If not, samples from seen and unseen classes would be forcefully aligned, leading to severe negative transfer [23]. Besides, diverse seen and unseen classes should be differentiated so that their recognition and semantic recovery would not be confused. In our

Algorithm 1 The complete procedure of AMS.

```
Training Procedure
for epoch=1 to $T_1$ do
    Fine-tune $f$, $g_s$ with Eq. (3).
end for
for epoch=1 to $T_2$ do
    Separate target samples from seen and unseen classes with Eq. (4).
    Pseudo-label target samples from unseen classes with Eq. (8).
    Train $f$, $g$, $\phi$ with Eq. (10).
end for
Use $h = g \circ f$ to classify target samples into seen and unseen classes.
Learn $W$ on target samples from seen classes with Eq. (11).
Inference Procedure
Use $h = g \circ f$ to classify target samples into seen and unseen classes.
Recovery attributes of seen and unseen target samples with $W$ and $\phi$, respectively.
```

proposed framework AMS (Fig. 1), we leverage a kind of angular-discriminative and robust representation to tackle both challenges in four phases. In phase I, we aim to learn discriminative representation of seen classes robust to anomalies, which is expected to benefit the seen/unseen separation in phase II. In phase III, we again resort to such representation to potently differentiate diverse seen and unseen classes. Lastly, in phase IV, we seek to further promote the interpretation of target seen classes by learning a visual-semantic re-projection. Details on each of the four phases are presented sequentially in section 3.3-3.6, and the overall procedure is shown in algorithm 1.

3.3 Discriminative and Robust Fine-tuning

As analyzed in section 3.2, the detection of unseen classes lies the foundation of SR-OSDA. Distance-based outlier detection [29,2] has proven effective for detecting visual outliers from unseen classes based on the distance between representations. However, representation learned via the classical combination of CNN and a Softmax layer will make a linear separable partition of the whole feature space [47], where intra-class distance can be much greater than inter-class distance. Though such property may well suffice supervised classification, it can pose a severe threat to the detection of unseen classes. As the entire feature space is partitioned for the seen classes, data from unseen classes could be easily classified into a seen class with high confidence. Besides, due to poor intra-class compactness, features of unseen classes could spread diffusely in the whole feature space, increasing the false positive rate of seen classes.

To address the above issue, we propose to fine-tune the feature extractor on source domain with an additive angular margin loss [8] instead of cross-entropy loss with an ordinary Softmax layer:

$$\mathcal{L}_{src}^{Arc} = \frac{1}{N_s}\sum_{i=1}^{N_s} -\log \frac{e^{s\cdot\cos(\theta_{i,y_s^i}+m)}}{e^{s\cdot\cos(\theta_{i,y_s^i}+m)} + \sum_{j=1,j\neq y_s^i}^{|C_s|} e^{s\cdot\cos\theta_{i,j}}}, \tag{1}$$

where m is the additive angular margin penalty, s is a re-scaling factor, and $\theta_{i,j} = \arccos \frac{W_j^\top f(x_s^i)}{\|W_j\| \cdot \|f(x_s^i)\|}$ is the angle between the i-th source feature extracted by the feature extractor $f : \mathcal{X} \to \mathbb{R}^d$ and the weights of the j-th neuron of the $|C_s|$-way classifier $g_s : \mathbb{R}^d \to \mathbb{R}^{|C_s|}$. Note that we fix $b_j = 0$ as in [8]. This objective guides the model to learn highly angular-discriminative representation of seen classes on a hypersphere.

However, Eq. (1) still leads to features that are approximately linearly separable and form a partition of the entire feature space, particularly when m is relatively small, which is not robust in the presence of anomalies. For instance, the decision boundary for class one in a binary case is $\cos(\theta_1 + m) > \cos\theta_2$, which can be rewritten as $\cos\theta_1 \cdot \cos m - \sin\theta_1 \cdot \sin m > \cos\theta_2$. When m is set to a small value such as 0.05 or 0.1, $\sin m \approx 0$, and therefore the decision boundary is approximately linearly separable: $(\frac{W_1^\top}{\|W_1\| \cdot \|f(x)\|} \cdot \cos m - \frac{W_2^\top}{\|W_2\| \cdot \|f(x)\|}) f(x) > 0$. Although a larger m brings in more non-linearity, it increases optimization difficulty and causes convergence problems. Thus, m is often set small in practice [8]. Inspired by [47] that regularizes training by distance to class prototypes, we propose a regularization term for the angle between the learned representation and corresponding classifier weight to learn intra-class compact and angular discriminative features with non-linear decision boundaries, making room for the future unseen classes:

$$\mathcal{L}_{src}^{reg} = \sum_{i=1}^{N_s} - \cos\theta_{i,y_s^i}. \tag{2}$$

Therefore, the overall objective of our fine-tuning phase is:

$$\min_{f,g_s} \mathcal{L}_{src} = \mathcal{L}_{src}^{Arc} + \mathcal{L}_{src}^{reg}. \tag{3}$$

3.4 Seen/Unseen Separation

Thanks to the angular regularization above, the decision boundary is less overfitting to seen classes, leaving out an open space, wherein unseen classes are tightly bounded and stay farther away from all seen classes. Besides, experimental evidence in metric-based few-shot learning shows that optimizing prototype-based metrics could facilitate class-discriminative features when generalizing to unseen classes [39,28]. Therefore, the features of unseen classes are expected to have better clustering properties (see Appendix for experimental evidence). In such a desirable feature space, we define the probability of a target sample x_t^i belonging to a seen class as $p(y_t^i = c | x_t^i) = \frac{e^{-d(x_t^i, \mu_s^c)}}{\sum_{c'} e^{-d(x_t^i, \mu_s^{c'})}}$, where μ_s^c is the feature centroid of the c-th known class in the source domain, and d is cosine distance. The target sample is then pseudo-labeled by $\hat{y}_t^i = \arg\max_{c'} p(y_t^i = c' | x_t^i)$ with confidence $p_t^i = \max_{c'} p(y_t^i = c' | x_t^i)$. Next, we adopt a class-wise threshold to re-pseudo-label each target sample x_t^i to ensure more balanced results:

$$\hat{y}_t^i = \begin{cases} \hat{y}_t^i & \text{if } p_t^i > \bar{p}_t^{\hat{y}_t^i} \\ |C_s| + 1 & \text{otherwise} \end{cases}, \tag{4}$$

where $\bar{p}_t^{\hat{y}_t^i}$ is the average confidence of target samples pseudo-labeled as class $\hat{y}_t^i$, and $|C_s|+1$ means a general unseen class. Afterwards, we use K-means to cluster the detected unknown samples into K clusters, and then use K-means again with the $|C_s|+K$ centroids as initialization to refine pseudo-labels for one time. To this end, all target samples are pseudo-labeled by $\hat{y} \in \{1,...,|C_s|+K\}$ with corresponding centroids $R = \{R_1,...,R_{|C_s|+K}\} \subset \mathbb{R}^{d\times(|C_s|+K)}$, and are separated into a seen set $\mathcal{D}_t^s$ and unseen set $\mathcal{D}_t^u$.

3.5 Alignment and Separation with Angular Margin

With all target samples pseudo-labeled, the focus now becomes aligning source and target features from the same seen class to boost recognition and separating the detected target features from unseen classes to avoid negative transfer. Meanwhile, the diversity of unseen classes should be preserved and their intrinsic difference should even be accentuated to facilitate discriminative and diverse semantic recovery results. To achieve this goal, [13] deployed center loss:

$$\mathcal{L}_c = \frac{1}{N_1}\sum_{i=1}^{N_1}\sum_{c=1}^{|C_s|+K}(\mathbb{I}_{[y^i=c]}d(f(x^i_{s/t}),R_c) - \frac{\mathbb{I}_{[y^i\neq c]}}{|C_s|+K-1}d(f(x^i_{s/t}),R_c)), \quad (5)$$

where $N_1 = N_s + N_t$, d is cosine distance, $\mathbb{I}$ is the indicator function, and y_i is ground-truth label and pseudo-label for source and target samples respectively. However, we argue that there are mainly two defects: (i) Eq. (5) requires features of different classes to be far away from each other, but does not demarcate a margin that specifies how far is enough, which could be less efficient in optimization, (ii) Eq. (5) only ensures the discriminativeness of visual representation, and thus could not guarantee the visual-semantic joint representation, which is more informative for recognition, is still discriminative enough.

To tackle the above issues, we again resort to additive angular margin loss with angular prototype regularization in the joint visual-semantic feature space:

$$\mathcal{L}_{Arc} = \frac{1}{N_2}\sum_{i=1}^{N_2} -\log\frac{e^{s\cdot\cos(\theta_{i,y_i}+m)}}{e^{s\cdot\cos(\theta_{i,y_i}+m)}+\sum_{j=1,j\neq y_i}^{|C_s|+K}e^{s\cdot\cos\theta_{i,j}}}. \quad (6)$$

Despite with a similar formulation to Eq. (1) at first glance, there are some substantial differences. First, instead of deploying a $(|C_s|+1)$-way classifier, we use a $(|C_s|+K)$-way classifier $g:\mathcal{R}^d\rightarrow\mathcal{R}^{C_s+K}$ for the more discriminative recognition of seen classes and diverse unseen classes. Second, the features now become the visual-semantic joint features. For labeled source samples $F_s = \{f(x_s^i)\oplus a_s^i, f(x_s^i)\oplus\tilde{a}_s^i\}$ for $x_s^i\in\mathcal{D}_s$, for target samples pseudo-labeled as seen classes $F_t^s = \{f(x_t^i)\oplus\hat{a}_t^i, f(x_t^i)\oplus\tilde{a}_t^i\}$ for $x_t^i\in\mathcal{D}_t^s$, and target samples pseudo-labeled as unseen classes $F_t^u = \{f(x_t^i)\oplus\tilde{a}_t^i\}$ for $x_t^i\in\mathcal{D}_t^u$, where a, $\hat{a}$ denotes ground-truth, pseudo-labeled attributes, $\tilde{a}$ denotes attributes predicted by the visual-semantic projector $\phi:\mathbb{R}^d\rightarrow\mathbb{R}^m$, and $\oplus$ denotes feature concatenation. Thus, the joint feature set is $F = F_s\cup F_t^s\cup F_t^u$ with cardinality N_2, and

$\theta_{i,j} = \arccos \frac{W_j^\top f_i}{\|W_j\| \cdot \|f_i\|}$. Similarly, the angular prototype regularization forms as:

$$\mathcal{L}_{reg} = \sum_{i=1}^{N_2} -\cos \theta_{i,y_i}. \tag{7}$$

It is worth noting that for training classifier g, the pseudo-labeling of target samples detected from unseen classes is non-trivial. After using K-means discussed in section 3.4, the target samples detected from unseen classes are pseudo-labeled as $\hat{y}^i_{unseen} \in \{|C_s| + 1, ..., |C_s| + K\}$, which seems plausible for training the classifier neurons corresponding to K unseen class clusters. However, since there is no guarantee that the same unseen cluster label generated by K-means in different iterations has the same semantic meaning, directly using such pseudo-labels could lead to inconsistent correspondence between samples and classifier neurons, increasing the difficulty of optimization. To address such problem, we propose a straightforward strategy termed **maximum-response correspondence (MRC)** to decide the pseudo-label $\mathring{y}^i_{unseen}$ of each x^i_{unseen} for training the classifier g:

$$\mathring{y}^i_{unseen} = \begin{cases} \hat{y}^i_{unseen} & \text{if current epoch=1} \\ \arg\max_{c \in C_u} p(y = c | x^i_{unseen}; \theta) & \text{otherwise} \end{cases}, \tag{8}$$

where $C_u = \{|C_s|+1, ..., |C_s|+K\}$ and θ denotes the parameters of the classifier.

With angular margin and regularization deployed, the learned representation is naturally more suitable for semantic recovery, since the joint visual-semantic representation is forced to be intra-class compact and inter-class separated, which means the input of the visual-semantic projection ϕ, i.e., visual features, as well as the output, i.e., semantic features, are both learned in a similar manner, savoring a mutual-reinforcing advantage. Hence, we can simply adopt the same binary cross-entropy objective as [13] but can observe much better results:

$$\min_{f,\phi} \mathcal{L}_A = \frac{1}{N_s + N_t^s} \sum_{x^i \in \mathcal{D}_s \cup \mathcal{D}_t^s} \mathcal{L}_{bce}(\hat{a}^i, a^i), \tag{9}$$

where N_s, N_t^s denotes the number of source and target samples detected from seen classes respectively, $\hat{a}^i = \sum_j W_{ij} \phi(f(x^j))$, and W is the propagator matrix [51,35] based on visual similarity.

Therefore,the objective of our training phase becomes:

$$\min_{f,g,\phi} \mathcal{L}_{AS} = \lambda_1 \mathcal{L}_c + \lambda_2 (\mathcal{L}_{Arc} + \mathcal{L}_{reg}) + \lambda_3 \mathcal{L}_A, \tag{10}$$

and we alternate between phase II and III in every training epoch to constantly reinforce model performance.

3.6 Attribute Re-Projection for Seen Classes

For visual-semantic projection, early works in ZSL typically resort to linear projection with regularization [36,14], while neural network-based projectors [50,26] are getting increasingly popular for their non-linearity and learnable representation. In SR-OSDA, the advantages of a neural-network-based projector trained end-to-end are conspicuous, as complementary signals back-propagated from both semantic and visual modalities can directly update the feature, making it more suitable for projecting to semantic space.

However, we note that in Eq. (9), the better generalizability to unseen classes of ϕ comes at the sacrifice of seen classes, since the propagated attribute vector is different from its original prediction. To address this issue, we propose to deploy an efficient attribute re-projection after training, in which we learn a regularized linear mapping from the visual representation of target seen classes to corresponding pseudo attributes:

$$\min_{W} \mathcal{L}(Wf(X_t^s), \hat{A}_t^s) + \varOmega(W), \tag{11}$$

where $f(X_t^s) \in \mathbb{R}^{d \times N_t^s}$ is the feature matrix of target samples recognized as from seen classes by g, $\hat{A}_t^s$ is the corresponding pseudo attribute matrix, $W \in \mathbb{R}^{m \times d}$ is the projection matrix, and $\varOmega$ is a certain regularization. For simplicity, we apply squared loss for Eq. (11) with a closed-form solution [36]. There are mainly two merits of such re-projection. First, X_t^s is based on the final classification result, which is expected to be more accurate. The projection matrix W can now exert its full power to project $f(X_t^s)$ to $\hat{A}_t^s$ as accurately as possible, and thus is expected to perform better than ϕ on seen classes. Second, unlike a neural-network-based projector that takes a long training time, W has a closed-form solution and can be solved instantly. Therefore, after the model is trained and used to separate target data into seen X_t^s and unseen X_t^u, we use W and ϕ to recover attributes for X_t^s and X_t^u respectively.

4 Experiments

4.1 Setup

Datasets. We evaluate our method on the two datasets curated by [13] for the SR-OSDA problem. (1) **D2AwA** is collected from the shared 17 classes of "real image" (R) and "painting" (P) domains of the DomainNet [32] dataset and the AwA2 [45] dataset. Following alphabetic order, the first 10 classes are chosen as shared seen classes, and the other 7 classes as unseen classes. (2) **I2AwA** [53] consists of the 3D2 (I) dataset and the AwA2 dataset. Since the 3D2 dataset only has 40 classes, it is used as the source domain, while the AwA2 dataset with 10 more classes serves as the target domain. We use the binary attributes of AwA2 as semantic description to evaluate our method, and only the ground-truth attributes of seen categories are available throughout training.

Evaluation metrics. Following the standard evaluation protocols in OSDA and ZSL, for the open-set recognition aspect, we evaluate the per-class average accuracy of seen (known) classes OS^*, the accuracy of unseen (unknown) class recognition $OS^\diamond$, and their harmonic mean $H_1 = \frac{2 \times OS^* \times OS^\diamond}{OS^* + OS^\diamond}$; for the semantic recovery aspect, we first determine if a sample is from seen or unseen classes

Table 1: Open-set domain adaptation accuracy (%) on D2AwA and I2AwA.

tasks	A→P			A→R			P→A			P→R			R→A			R→P			I→A		
method	OS^*	$OS^\diamond$	H_1	OS^*	$OS^\diamond$	H_1	OS^*	$OS^\diamond$	H_1	OS^*	$OS^\diamond$	H_1	OS^*	$OS^\diamond$	H_1	OS^*	$OS^\diamond$	H_1	OS^*	$OS^\diamond$	H_1
Source-only	63.5	0.0	0.0	86.1	0.0	0.0	92.5	0.0	0.0	86.8	0.0	0.0	96.2	0.0	0.0	68.3	0.0	0.0	76.2	0.0	0.0
OSBP [38]†	49.6	10.8	17.7	74.2	13.6	23.0	76.0	9.1	16.3	63.3	6.9	12.4	90.1	13.7	23.8	55.9	10.6	17.8	67.6	7.5	13.5
STA [25]†	60.1	33.0	42.6	85.5	10.8	19.2	90.2	5.7	10.7	82.8	7.4	13.58	88.5	7.2	13.3	66.9	13.5	22.5	51.5	45.5	48.3
AOD [10]†	50.7	9.5	16.0	78.4	12.7	21.9	80.3	5.1	9.6	79.7	5.3	9.9	92.0	12.8	22.5	61.2	9.6	16.6	75.2	6.3	11.6
BCA [12]†	35.8	86.4	50.6	73.4	87.5	**79.8**	75.7	91.9	82.9	63.3	85.3	69.5	90.6	92.9	91.7	55.1	77.7	64.5	2.3	53.9	4.3
SR-OSDA [13]†	42.6	83.1	56.3	80.4	76.4	78.3	79.5	96.6	**87.2**	77.9	88.1	82.7	91.4	93.8	92.6	53.2	85.3	65.6	68.3	70.2	69.2
AMS (ours)	48.1	79.5	**59.9**	83.4	76.3	79.6	79.3	96.5	87.0	82.4	91.7	**86.8**	94.7	92.6	**93.7**	63.0	77.0	**69.3**	76.3	75.6	**75.9**

† Cited from [13]. † Reproduced with official codes. † Reproduced by us.
* Emboldened figures: the best balanced performance H_1.

Table 2: Semantic recovery accuracy (%) on D2AwA and I2AwA.

tasks	A→P			A→R			P→A			P→R			R→A			R→P			I→A		
method	S	U	H_2	S	U	H_2	S	U	H_2	S	U	H_2	S	U	H_2	S	U	H_2	S	U	H_2
Source-only	67.6	0.0	0.0	87.6	0.0	0.0	91.3	0.0	0.0	85.3	0.0	0.0	94.1	0.0	0.0	71.1	0.0	0.0	77.2	0.3	0.7
ABP* [52]†	68.1	0.0	0.0	87.9	0.0	0.0	91.7	0.0	0.0	83.6	0.0	0.0	94.4	0.0	0.0	70.0	0.0	0.0	79.8	0.0	0.0
TF-VAE* [30]†	70.4	0.0	0.0	88.4	0.0	0.0	85.1	0.0	0.0	79.6	0.0	0.0	96.4	0.0	0.0	72.5	0.0	0.0	62.8	0.0	0.0
ABP [52]†	64.5	6.4	11.7	86.0	5.9	11.1	84.0	24.4	37.8	81.3	12.7	21.9	93.8	16.2	27.6	67.6	7.9	14.1	78.0	13.4	22.9
TF-VAE [30]†	59.7	12.8	21.0	77.9	16.4	27.1	35.1	35.6	35.3	34.8	32.7	33.7	68.5	36.1	47.3	50.7	21.0	29.7	37.7	20.0	26.2
SR-OSDA [13]†	42.7	20.2	27.4	80.3	34.3	48.0	77.5	50.9	61.4	77.7	45.6	57.4	90.0	49.2	63.6	52.9	24.0	32.9	59.4	27.8	37.8
AMS (ours)	48.1	24.8	**32.7**	83.5	45.9	**59.1**	79.8	64.0	**71.0**	82.8	58.1	**68.3**	94.9	59.1	**72.8**	63.2	25.2	**36.0**	74.8	28.3	**41.0**

† Cited from [13]. † Reproduced by us. * Emboldened figures: the best balanced performance H_2.

based on the recognition result, and then make inference based on the matching of recovered semantic attributes with class semantic attribute prototypes in the corresponding range (C_s or C_u) by cosine similarity, and evaluate the per-class average accuracy of seen class recovery S, unseen class recovery U, and their harmonic mean $H_2 = \frac{2\times S\times U}{S+U}$.

Implementation details. We use ResNet-50 [11] pre-trained on ImageNet [37] as backbone. For fair comparison, we adopt the same network architecture for feature encoder, attribute projector, and classifier as [13]. If not specified, we set K to the ground-truth unseen cluster number, as we experimentally notice the performance is not sensitive to its value in a certain range. We set $m = 0.05$, $s = 30$, $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, and $\lambda_3 = 1$ in all experiments. Since [13] has not released code at the time of our work, we report our reproduced results in this paper. Please refer to the Appendix for more details as well as openness analysis, qualitative study, and parameter analysis.

4.2 Experimental Results

Open-set domain adaptation evaluation. From results in table 1 we can see that existing pure OSDA methods tend to underperform in all tasks because they often learn to automatically separate samples into seen and unseen classes via adversarial learning, which is unstable and could either accept most target samples into seen classes or reject them into unseen classes, leading to highly imbalanced performance (high OS^* or $OS^\diamond$, but low H_1). Besides, OSDA methods are not designed to work with semantic information and cannot leverage the complementary information in semantic attributes. In addition, our proposed AMS notably outperforms the SR-OSDA baseline: our accuracy on seen class (OS^*)

Table 3: The evolution path of AMS (in P→R). Margin: additive angular margin. Reg: angular regularization. Multi-unseen: $|C_s + K|$-way classifier instead of $|C_s + 1|$-way. MRC: maximum-response correspondence strategy. Re-projection: regularized linear projection on detected target samples from seen classes.

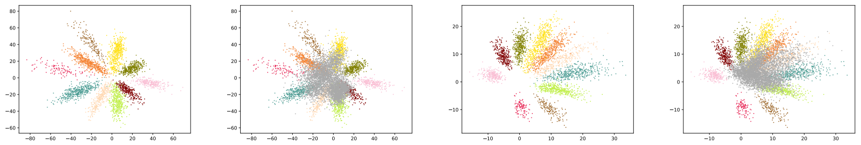
modules	base [13]									AMS
fine-tuning margin		✓	✓	✓	✓	✓	✓	✓	✓	✓
fine-tuning reg			✓	✓	✓	✓	✓	✓	✓	✓
training margin				✓	✓	✓	✓	✓	✓	✓
training reg					✓		✓		✓	✓
training multi-unseen						✓	✓	✓	✓	✓
training MRC								✓	✓	✓
post-training re-projection										✓
OS^*(%)	77.9	77.8	78.3	82.6	83.7	85.7	85.8	81.8	82.4	82.4
$OS^\diamond$(%)	88.1	90.3	90.2	90.4	89.5	82.9	81.5	91.4	91.7	91.7
H_1(%)	82.7	83.6	83.2	86.3	86.5	84.3	83.6	86.3	86.8	86.8
S(%)	77.7	77.2	77.6	82.6	83.3	84.9	85.3	81.1	81.7	82.8
U(%)	45.6	50.9	50.7	47.3	41.6	47.3	40.5	56.8	58.1	58.1
H_2(%)	57.4	61.3	61.3	60.1	55.4	60.7	54.9	66.8	67.9	68.3

is higher in *6 out of 7* tasks with an average improvement of *4.8%*, accuracy on unseen class ($OS^\diamond$) is higher or comparable in *4 out of 7* tasks with a notable improvement of 5.4% in I→A, and harmonic mean H_1 is higher or comparable in *all 7* tasks with an average increment of 2.9%, suggesting that we achieve *a balanced improvement* on seen and unseen recognition accuracy.

Semantic recovery evaluation. We compare our method with several ZSL methods as well as the recently-proposed SR-OSDA baseline on the task of semantic recovery in table 2. ABP [52] and TF-VAE [30] are both GZSL methods that need access to semantic attributes of unseen classes, and therefore violate the SR-OSDA setting. We also report their results using only semantic attributes of seen classes as ABP* and TF-VAE*. From table 2 we can have two observations. Firstly, the semantic recovery capability of semantic-transductive GZSL methods largely relies upon the prior knowledge of unseen classes, whereas even when given such prior knowledge they still tend to not behave well because they cannot cope with domain gap. Secondly, our proposed AMS significantly outperforms the existing SR-OSDA baseline on all metrics, with a notable average increment of *6.6%*, *7.6%*, and *7.4%* in S, U, and H_2 respectively.

4.3 Ablation Study

Fine-tuning phase. As is shown in table 3, fine-tuning with additive margin loss notably increases the final $OS^\diamond$ and U of *unseen* classes by 2.2% and 5.3%, respectively. This result verifies that the model can benefit from the highly discriminative representation that is initially learned on the seen classes in source domain, even when such properties of representation are no longer purposefully emphasized during the entire next training phase. Further, when angular regularization is applied in fine-tuning, OS^* and S of *seen* classes are further improved by 0.5% and 0.4% respectively, which verifies that robust representation of seen classes is beneficial to their recognition and recovery in presence of outliers.



(a) (b) (c) (d)

Fig. 2: Visualization of representation learned with/without the proposed fine-tuning phase. Points of bright color are from seen classes in the source domain. Gray points are from unseen classes in the target domain. Best viewed in color.

Training phase. As is shown in table 3, applying angular margin loss in training brings salient improvements on all evaluation metrics except for a decrease on U. This phenomenon is because to this end the classifier is set to $|C_s + 1|$-way, which treats different unseen classes as one class. As a result, though recognition accuracy on seen classes and the overall unseen class benefits from the learned discriminative representation, the diverse unseen classes are forced to be aligned, leading to confused semantic recovery. Besides, if angular regularization is added at this time, there is a further decrease in U, which is caused by further pulling different unseen classes closer.

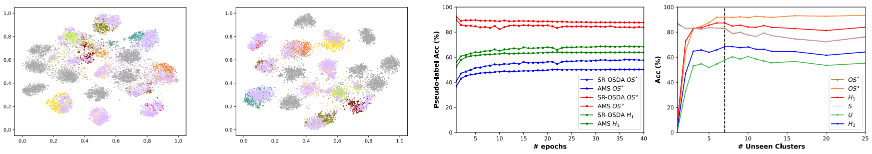
With such observation, we now turn the classifier to $|C_s + K|$-way to cope with diverse unseen classes, but still cannot see a notable improvement. This is caused by the inconsistent correspondence between clusters of unseen classes and classifier neurons in different training epochs, which hinders the knowledge learning of unseen classes. Therefore, it is not until the maximum-response correspondence (MRC) mechanism is adopted that the power of multiple unseen clusters can be truly exerted, bringing notable improvements on all metrics, and particularly, semantic recovery accuracy U of unseen classes is increased by a whopping 12.5%.

To this end, we can conclude that in the training phase, for semantic recovery of unseen classes U, angular regularization is beneficial only when the classifier has *multiple* channels for unseen classes, while such classifier should work under the *MRC* mechanism.

Post-training re-projection. Results in the last two columns in table 3 verify the effectiveness of learning a re-projection on detected target seen classes after training, bringing an extra 1.1% increment on seen class recovery S, which is non-trivial for accuracy averaged over 10 classes with more than 10,000 images in total. Note that the only difference between the last two columns lies in their S results, since the re-projection is only learned on target samples recognized as from seen class after the training phase, and therefore will not have any effect on unseen classes as they are now completely separated.

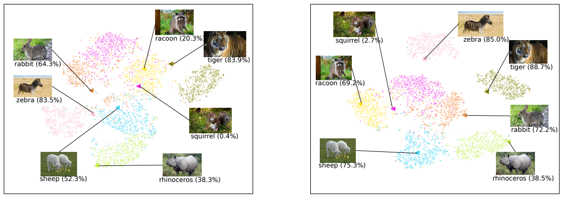
4.4 Visual Study of Representation

Visual representation learned in the fine-tuning phase. To better observe the property of representation learned in fine-tuning phase, we set the dimen-



(a) (b) (c) (d)

Fig. 3: Visualization of representation and relevant studies of training phase. (a),(b): Lavender and gray points are from seen and unseen classes in the target domain respectively. Points in other colors are from different seen classes in the source domain. (c): Accuracy of pseudo-labels determined by K-means (in task R→P) during training. (d): Final accuracy of recognition and semantic recovery using different number of clusters for unseen classes. Best viewed in color.



(a) (b)

Fig. 4: Visualization of semantic representation (projected attributes). Round points in different colors are representations of target samples from different unseen classes, while triangles are their ground-truth prototypes accompanied by a typical image. The percentage in parentheses reports accuracy of semantic recovery. Best viewed in color.

sion of visual features to 2 and plot them in Fig. 2. Fig. 2 (a) and (c) visualize the representation of *seen* classes learned with ordinary Softmax and with our proposed fine-tuning phase in task P→R, respectively. We can see that the former is linearly separable, making a full partition of the entire feature space. In contrast, after our fine-tuning, there is an open space left out at the center of the feature space, and the decision boundary is no longer linear. As the result, when deployed on the target domain, as is shown in Fig. 2 (b),(c), in the former case, samples from *unseen* classes are misclassified into almost all *seen* classes, while in the latter case, part of samples from *unseen* classes are rejected into the vacant space at the center, and *seen* classes are less affected.

Visual representation learned in the training phase. Fig. 3 (a),(b) show the T-SNE [27] visualization of representation learned by the SR-OSDA framework proposed in [13] and our AMS in task P→R, respectively. We can observe that feature clusters of both *seen* and *unseen* classes are more compact, and different clusters are better separated in AMS, thanks to the angular margin and regularization. We also show the accuracy of pseudo-labels determined by K-means clustering in task R→P in Fig. 3 (c) and can see that AMS achieves a better balance between seen and unseen classes, i.e., a higher H_1, which underpins the better performance of the later cross-domain alignment and recognition.

Semantic representation learned in the training phase. Fig. 4 (a),(b) show the T-SNE [27] visualization of semantic representation (projected attributes) of *unseen* classes in task P→R. Compared with [13], our semantic representation mainly has two merits. First, our semantic representation is more compactly clustered around corresponding semantic prototypes. For instance, our representations of zebra and tiger better enclose their prototypes, which boosts their recovery accuracy by 1.5% and 4.8%. Second, our semantic representations of different classes are more separated. For example, our representations of sheep and racoon are farther away from the prototypes of other classes, which increases their recovery accuracy by 23% and 48.9%. These phenomena verify the functionality of angular margin and regularization for semantic recovery.

How does the number of clusters affect learning? Fig. 3 (d) shows the performance of AMS in task P→R when the cluster number K varies from 1 to up to 25. It can be observed that when K is very small, $OS^{\diamond}$ and U are low since the model nearly completely ignores the intrinsic diversity of unseen classes. As K grows, $OS^{\diamond}$ and U rise rapidly, resulting in more balanced H_1 and H_2. When K surpasses the ground truth, as is to the right of the dashed line in Fig. 3 (d), performance on all metrics is still stable, even when K reaches more than three times that of the ground truth. This result verifies the robustness of AMS and also suggests we choose a relatively large cluster number in real-world deployment.

5 Conclusion

In this paper, we present a novel framework termed AMS for the practical semantic recovery open-set domain adaptation challenge. At the core of our conception is the widely acknowledged significance of representation learning for visual tasks. To learn discriminative visual representation robust to outliers, AMS first fine-tunes a pre-trained model on the seen classes in the source domain with an additive angular margin and angular regularization. Grounded by such learned initial representation, AMS performs cross-domain alignment on seen classes, separates unseen classes from seen ones, and accentuates the intrinsic diversity of unseen classes by resorting to additive angular margin and angular regularization again on the joint visual-semantic representation of the target domain. Further, AMS adopts an efficient post-training re-projection to boost semantic recovery of target seen classes without hurting that of unseen classes. Extensive quantitative experiments as well as various visual studies verify that AMS achieves competitive and even state-of-the-art performance.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 62176042 and 62073059, in part by CCF-Baidu Open Fund (NO. 2021PP15002000), in part by CCF-Tencent Open Fund (NO. RAGR20210107), and in part by Guangdong Basic and Applied Basic Research Foundation (No. 2021B1515140013).

References

1. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
2. Bucci, S., Borlino, F.C., Caputo, B., Tommasi, T.: Distance-based hyperspherical classification for multi-source open-set domain adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1119–1128 (2022)
3. Bucci, S., Loghmani, M.R., Tommasi, T.: On the effectiveness of image rotation for open set domain adaptation. In: European Conference on Computer Vision. pp. 422–438. Springer (2020)
4. Chao, W.L., Changpinyo, S., Gong, B., Sha, F.: An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: European conference on computer vision. pp. 52–68. Springer (2016)
5. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3339–3348 (2018)
6. Choi, J., Sharma, G., Schulter, S., Huang, J.B.: Shuffle and attend: Video domain adaptation. In: European Conference on Computer Vision. pp. 678–695. Springer (2020)
7. Csurka, G.: Domain adaptation for visual applications: A comprehensive survey. arXiv preprint arXiv:1702.05374 (2017)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
9. Du, Z., Li, J., Lu, K., Zhu, L., Huang, Z.: Learning transferrable and interpretable representations for domain generalization. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3340–3349 (2021)
10. Feng, Q., Kang, G., Fan, H., Yang, Y.: Attract or distract: Exploit the margin of open set. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7990–7999 (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Jing, M., Li, J., Zhu, L., Ding, Z., Lu, K., Yang, Y.: Balanced open set domain adaptation via centroid alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 8013–8020 (2021)
13. Jing, T., Liu, H., Ding, Z.: Towards novel target discovery through open-set domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9322–9331 (2021)
14. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3174–3183 (2017)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
16. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 951–958. IEEE (2009)

17. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
18. Li, J., Chen, E., Ding, Z., Zhu, L., Lu, K., Shen, H.T.: Maximum density divergence for domain adaptation. IEEE transactions on pattern analysis and machine intelligence **43**(11), 3918–3930 (2020)
19. Li, J., Du, Z., Zhu, L., Ding, Z., Lu, K., Shen, H.T.: Divergence-agnostic unsupervised domain adaptation by adversarial attacks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
20. Li, J., Jing, M., Lu, K., Ding, Z., Zhu, L., Huang, Z.: Leveraging the invariant side of generative zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7402–7411 (2019)
21. Li, J., Jing, M., Zhu, L., Ding, Z., Lu, K., Yang, Y.: Learning modality-invariant latent representations for generalized zero-shot learning. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1348–1356 (2020)
22. Li, J., Lu, K., Huang, Z., Zhu, L., Shen, H.T.: Transfer independently together: A generalized framework for domain adaptation. IEEE transactions on cybernetics **49**(6), 2144–2155 (2018)
23. Li, X., Li, J., Zhu, L., Wang, G., Huang, Z.: Imbalanced source-free domain adaptation. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3330–3339 (2021)
24. Li, Y., Wang, D., Hu, H., Lin, Y., Zhuang, Y.: Zero-shot recognition using dual visual-semantic mapping paths. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3279–3287 (2017)
25. Liu, H., Cao, Z., Long, M., Wang, J., 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)
26. Liu, S., Long, M., Wang, J., Jordan, M.I.: Generalized zero-shot learning with deep calibration network. Advances in Neural Information Processing Systems **31** (2018)
27. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
28. Mensink, T., Verbeek, J., Perronnin, F., Csurka, G.: Distance-based image classification: Generalizing to new classes at near-zero cost. IEEE transactions on pattern analysis and machine intelligence **35**(11), 2624–2637 (2013)
29. Miller, D., Sunderhauf, N., Milford, M., Dayoub, F.: Class anchor clustering: A loss for distance-based open set recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3570–3578 (2021)
30. Narayan, S., Gupta, A., Khan, F.S., Snoek, C.G., Shao, L.: Latent embedding feedback and discriminative features for zero-shot classification. In: European Conference on Computer Vision. pp. 479–495. Springer (2020)
31. Panareda Busto, P., Gall, J.: Open set domain adaptation. In: Proceedings of the IEEE international conference on computer vision. pp. 754–763 (2017)
32. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1406–1415 (2019)
33. Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C.P., Wang, X.Z.: A review of generalized zero-shot learning methods. arXiv preprint arXiv:2011.08641 (2020)
34. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

35. Rodríguez, P., Laradji, I., Drouin, A., Lacoste, A.: Embedding propagation: Smoother manifold for few-shot classification. In: European Conference on Computer Vision. pp. 121–138. Springer (2020)
36. Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: International conference on machine learning. pp. 2152–2161. PMLR (2015)
37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
38. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 153–168 (2018)
39. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)
40. Song, J., Shen, C., Yang, Y., Liu, Y., Song, M.: Transductive unbiased embedding for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1024–1033 (2018)
41. Su, H., Li, J., Chen, Z., Zhu, L., Lu, K.: Distinguishing unseen from seen for generalized zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7885–7894 (2022)
42. Wang, M., Deng, W.: Deep visual domain adaptation: A survey. Neurocomputing **312**, 135–153 (2018)
43. Wang, X., Ye, Y., Gupta, A.: Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6857–6866 (2018)
44. Wang, Z., Gou, Y., Li, J., Zhang, Y., Yang, Y.: Region semantically aligned network for zero-shot learning. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 2080–2090 (2021)
45. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learninga comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence **41**(9), 2251–2265 (2018)
46. Xu, X., Shen, F., Yang, Y., Zhang, D., Tao Shen, H., Song, J.: Matrix tri-factorization with manifold regularizations for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3798–3807 (2017)
47. Yang, H.M., Zhang, X.Y., Yin, F., Liu, C.L.: Robust classification with convolutional prototype learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3474–3482 (2018)
48. Ye, M., Guo, Y.: Zero-shot classification with discriminative semantic representation learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7140–7148 (2017)
49. You, F., Li, J., Zhu, L., Chen, Z., Huang, Z.: Domain adaptive semantic segmentation without source data. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3293–3302 (2021)
50. Zhang, L., Xiang, T., Gong, S.: Learning a deep embedding model for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2021–2030 (2017)
51. Zhou, D., Bousquet, O., Lal, T., Weston, J., Schölkopf, B.: Learning with local and global consistency. Advances in neural information processing systems **16** (2003)

52. Zhu, Y., Xie, J., Liu, B., Elgammal, A.: Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9844–9854 (2019)
53. Zhuo, J., Wang, S., Cui, S., Huang, Q.: Unsupervised open domain recognition by semantic discrepancy minimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 750–759 (2019)