International Journal of Computer Vision (2021) 129:1893–1909
https://doi.org/10.1007/s11263-021-01451-1

Visual Structure Constraint for Transductive Zero-Shot Learning in the Wild

Ziyu Wan[1] · Dongdong Chen[2] · Jing Liao[1]

Received: 21 December 2019 / Accepted: 24 February 2021 / Published online: 19 April 2021

Abstract
To recognize objects of the unseen classes, most existing Zero-Shot Learning(ZSL) methods first learn a compatible projection function between the common semantic space and the visual space based on the data of source seen classes, then directly apply it to the target unseen classes. However, for data in the wild, distributions between the source and target domain might not match well, thus causing the well-known domain shift problem. Based on the observation that visual features of test instances can be separated into different clusters, we propose a new visual structure constraint on class centers for transductive ZSL, to improve the generality of the projection function (*i.e.*alleviate the above domain shift problem). Specifically, three different strategies (symmetric Chamfer-distance, Bipartite matching distance, and Wasserstein distance) are adopted to align the projected unseen semantic centers and visual cluster centers of test instances. We also propose two new training strategies to handle the data in the wild, where many unrelated images in the test dataset may exist. This realistic setting has never been considered in previous methods. Extensive experiments demonstrate that the proposed visual structure constraint brings substantial performance gain consistently and the new training strategies make it generalize well for data in the wild. The source code is available at https://github.com/raywzy/VSC.

Keywords Computer vision · Zero-shot learning · Visual structure constraint

1 Introduction

Relying on massive labeled training datasets, significant progress has been made for image recognition in the past years (Simonyan and Zisserman 2014; Szegedy et al. 2015; He et al. 2016). However, it is unrealistic to label all the object classes, thus making these supervised learning methods struggle to recognize objects which are unseen during training. By contrast, Zero-Shot Learning (ZSL) (Norouzi et al. 2014; Zhang and Saligrama 2016b; Zhu et al. 2019; Long et al. 2017) only requires labeled images of seen classes (source domain), and are capable of recognizing images of unseen classes (target domain). The seen and unseen domains often share a common semantic space, which defines how unseen classes are semantically related to seen classes. The most popular semantic space is based on semantic attributes, where each seen or unseen class is represented by an attribute vector. Besides the semantic space, images of the source and target domains are also related and represented in a visual feature space.

To associate the semantic space and the visual space, existing methods often rely on the source domain data to learn a compatible projection function to map one space to the other, or two compatible projection functions to map both spaces into one common embedding space. During test time, to recognize an image in the target domain, semantic vectors of all unseen classes and the visual feature of this image would be projected into the embedding space using the learned function, then nearest neighbor (NN) search will be performed to find the best match class. However, due to the existence of the distribution difference between the source and target domains, especially for in-the-wild data, the learned projec-

Communicated by Mei Chen.

✉ Jing Liao
jingliao@cityu.edu.hk

Ziyu Wan
ziyuwan2-c@my.cityu.edu.hk

Dongdong Chen
cddlyf@gmail.com

[1] City University of Hong Kong, Kowloon, Hong Kong

[2] Microsoft Cloud+AI, Lexington, USA

tion function often suffers from the well-known **domain shift** problem.

To compensate for this domain gap, transductive zero-shot learning (Fu et al. 2015) assumes that the semantic information (*e.g.*attributes) of unseen classes and visual features of all test images are known in advance. Different ways like domain adaption (Kodirov et al. 2015) and label propagation (Zhu and Ghahramani 2002) are well investigated to better leverage this extra information. Recently, Zhang and Saligrama (2016b) find that visual features of unseen target instances can be separated into different clusters even though their labels are unknown as shown in Fig. 1. By incorporating this prior as a regularization term, a better label assignment matrix can be solved with a non-convex optimization procedure. However, their method still has three main limitations: 1) This visual structure prior is not used to learn a better projection, which directly limits the upper bound of the final performance. 2) They model the ZSL problem as a less-scalable batch mode, which requires re-optimization when adding new test data. 3) Like most previous transductive ZSL methods, they have not considered a realistic scenario where many unrelated images may exist in the test dataset and make the above prior invalid.

Considering the first problem, we model the above visual structure prior as a new constraint to learn a better projection function rather than use the pre-defined one. In this paper, we adopt the visual space as the embedding space and project the semantic space into it. To learn the projection function, we not only use the projection constraint of the source domain data as Zhang et al. (2017) but also impose the aforementioned visual structure constraint of the target domain data. Specifically, during training, we first project all the unseen semantic classes into the visual space, then consider three different strategies ("Chamfer-distance based", "Bipartite matching based" and "Wasserstein-distance based") to align the projected unseen semantic centers and the visual centers. However, due to the lack of test instance labels in the ZSL setting, we approximate these real visual centers with some unsupervised clustering algorithms (*e.g.*K-Means). Need to note that in our method, we directly apply the learned projection function to the online-mode testing, which is more friendly to real applications when compared to the batch mode in Zhang and Saligrama (2016b).

The third problem is common for data in the wild, where many unrelated images, which belong to neither seen nor unseen classes, often exist in the target domain. Thus using current unsupervised clustering algorithms directly on the whole test dataset will generate invalid visual centers, and misguide the learning of the projection functions. To overcome this problem, we further propose a new training strategies that first filters out the highly unrelated images and then uses the remaining ones to impose the proposed visual constraint. The filter is based on the distance from each image to its closest visual center. Considering the initial projection function is often deviated more than the later refined one, we further propose another progressive training strategy to gradually refine the center and increase the threshold in a class-specific manner. To the best of our knowledge, we are the first to consider this realistic transductive ZSL configuration with unrelated test images.

We demonstrate the effectiveness of the proposed visual structure constraint on many different widely-used datasets. Extensive experiments on both small datasets like AwA2 and large-scale dataset like Imagenet all show that the proposed visual structure constraint can consistently bring substantial performance gain and achieve state-of-the-art results.

To summarize, our contributions are three-fold as below:

- We have proposed three different types of visual structure constraint for the projection learning of transductive ZSL to alleviate its domain shift problem.
- We introduce a realistic transductive ZSL configuration where many unrelated images exist in the test dataset and propose two new training strategies to make our method work for it.
- Experiments demonstrate that the proposed visual structure constraint can bring substantial performance gain consistently and achieve state-of-the-art results.

This paper is extended from our preliminary conference version (Wan et al. 2019) in four ways. First, we propose a new progressive training strategy for the realistic setting of data in the wild. Second, we carry out extensive experiments on the large-scale dataset like ImageNet and push current state-of-the-art performance by a large margin. This further demonstrates the scalability and effectiveness of our method. Third, we extend the wasserstein-based visual structure constraint to its instance-based counterpart and show its robustness even in some extremely challenging cases. Fourth, we present deeper and more detailed analysis and discussion of the proposed method.

2 Related Work

Zero-shot Learning and Semantic Spaces. Though deep supervised learning has gained enormous success for the image recognition task (Simonyan and Zisserman 2014; Szegedy et al. 2015; He et al. 2016), it relies on large-scale human annotations and cannot generalize to new unseen classes. Zero-shot learning bridges the gap between training seen classes and testing unseen classes via different kinds of semantic spaces. Among them, the most popular and effective one is the attribute-based semantic space (Farhadi et al. 2009; Akata et al. 2013; Lampert et al. 2009). The attributes are often designed by experts to represent the class-specific properties, which are demonstrated to be reliable and effective. To incorporate more attributes for fine-grained recognition tasks, the text description-based semantic space is proposed in Reed et al. (2016), Zhang et al. (2017), Elhoseiny et al. (2013), Zhu et al. (2018), which provides a more natural language interface. Compared to these two labor-intensive types of methods, word vector-based methods (Frome et al. 2013; Miller 1995; Norouzi et al. 2014; Wang et al. 2018) can learn the semantic space from large text corpus automatically and save much human labor. However, they often suffer from visual-semantic discrepancy problem and achieve inferior performance.

Embedding Spaces. To relate the visual features of test images and semantic attributes of unseen classes, three different embedding spaces are often used by existing zero-shot learning methods: the original semantic space, the original visual space, and the newly-learned common intermediate embedding space. Specifically, they often learn a projection function from the visual space to the semantic space (Lampert et al. 2014; Reed et al. 2016; Frome et al. 2013; Annadani and Biswas 2018) or from the semantic space to the visual space (Kodirov et al. 2015; Shigeto et al. 2015; Zhang et al. 2017) in the first two cases, or learn two projection functions from semantic and visual space to the common embedding space (Changpinyo et al. 2016; Lu 2016; Zhang and Saligrama 2016a) respectively, which can be modeled as a regression or ranking problem solved by conventional methods or deep neural networks. In our method, we also use the visual space as the embedding space, because it is demonstrated helpful in alleviating the hubness problem (Radovanović et al. 2010) in Zhang et al. (2017). More importantly, our structure constraint is based on the separability of visual features of unseen classes.

GAN-based Zero-shot Learning. With the recent progress of generative adversarial networks(GAN) (Goodfellow et al. 2014), a series of GAN-based zero-learning methods (Xian et al. 2018; Felix et al. 2018; Xian et al. 2019; Huang et al. 2019; Li et al. 2019) have been proposed to solve the domain shift problem. For example, f-CLSWGAN (Xian et al. 2018) directly employs the conditional Wasserstein GAN (Arjovsky et al. 2017) to generate the sample features of unseen classes based on the class attributes, which could be utilized for consequent supervised learning. To ensure the synthetic visual features can recover their semantic features back, Felix et al. (2018) further add a multi-modal cycle-consistent constraint into GAN-based ZSL. In LisGAN (Li et al. 2019), a novel sample strategy is proposed to improve the ZSL performance. To achieve any-shot learning, Xian et al. (2019) design a f-VAEGAN-D2 framework which combines both GAN and VAE (Kingma and Welling 2013).

Although the above-mentioned GAN-based ZSL methods can achieve promising results in both conventional ZSL and generalized ZSL, they have some drawbacks compared to projection-based methods. Firstly, like other generative models, the training instability also exists in GAN-based ZSL. Besides, enough samples are needed to narrow down the gap between the generated distributions and the real feature distributions. When the training samples for each seen class are not enough, the generator usually could not capture the underlying distribution well. In contrast, our method could alleviate this problem well since the training only relies on the visual center. Last but not least, GAN-based ZSL could not be directly applied under the proposed realistic transductive setting because of the overlap between noise distribution and real distribution.

Domain Shift and Transductive Zero-shot Learning. Since the target unseen classes are disjointed to source seen classes, their underlying data distribution might be very different. In such cases, if we only learn the projection functions based on the source domain data, the learned projection functions will be biased and incur a serious performance gap when directly applied to the target domain. This problem is first witnessed by Fu et al. (2015) and alleviated by a new transductive zero-shot learning setting. Under this setting, the unseen target domain data is leveraged in the learning stage to improve the generalization ability and different strategies have been proposed. For example, a multi-view semantic space alignment process is used to correlate different semantic views and the low-level feature view by projecting them onto a common latent space learned using multi-view canonical correlation analysis in Fu et al. (2015). And unsupervised domain adaption is utilized in Kodirov et al. (2015) by proposing a novel regularized sparse coding framework. Our method also belongs to transductive approaches, and the proposed visual structure constraint is inspired by Zhang and Saligrama (2016b), but we have addressed their aforementioned drawbacks and improved the performance significantly. To the best of our knowledge, we are the first that utilizes the structure of visual space to constrain the projection function learning in transductive ZSL.

3 Method

Problem Definition. In ZSL setting, we have N_s source labeled samples $\mathcal{D}_s \equiv \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, where x_i^s is an image and $y_i^s \in \mathcal{Y}_s = \{1, \dots, S\}$ is the corresponding label within total S source classes. We are also given N_u unlabeled target samples $\mathcal{D}_u \equiv \{x_i^u\}_{i=1}^{N_u}$ that are from target classes $\mathcal{Y}_u = \{S+1, \dots, S+U\}$. According to the definition of ZSL, there is no overlap between source seen classes $\mathcal{Y}_s$ and target unseen classes $\mathcal{Y}_u$, *i.e.*$\mathcal{Y}_s \cap \mathcal{Y}_u = \emptyset$. But they are associated in a common semantic space, which is the knowledge bridge between the source and target domain. As explained before, we adopt semantic attribute space here, where each class $z \in \mathcal{Y}_s \cup \mathcal{Y}_u$ is represented with a pre-defined auxiliary

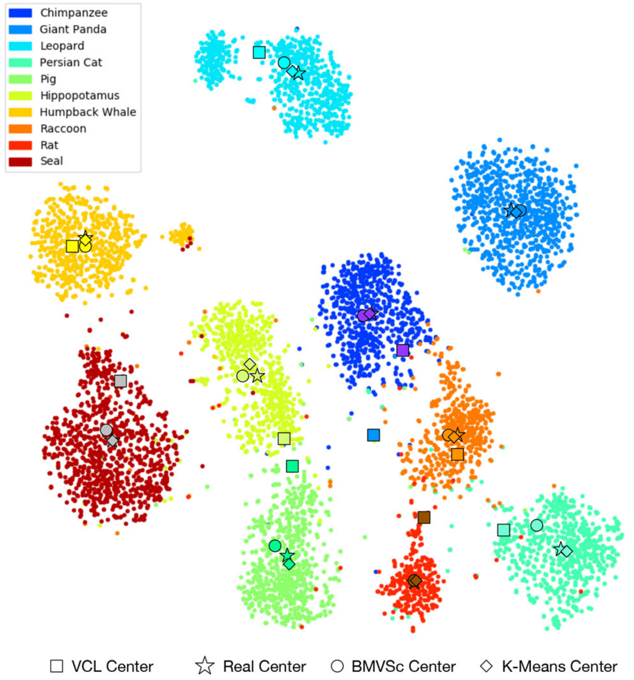

Fig. 1 Visualization of CNN feature distribution of 10 target unseen classes on AwA2 dataset using t-SNE, which can be clearly clustered into several real centers (stars). Squares (VCL) are synthetic centers projected by the projection function learned only from source domain data. By incorporating our visual structure constraint, our method (BMVSc) can help to learn better projection function and the generated synthetic semantic centers would be much closer to the real visual centers

attribute vector $a_z \in \mathcal{A}$. The goal of ZSL is to predict the label $y_i^u \in \mathcal{Y}_u$ given x_i^u with no labeled training data.

Besides the semantic representations, images of the source and target domains are also represented with their corresponding features in a common visual space. To relate these two spaces, projection functions are often learned to project these two spaces into a common embedding space. Following Zhang et al. (2017), we directly use the visual space as the embedding space, in which case only one projection function is needed. The key problem then becomes how to learn a better and generalized projection function.

Motivation. Our method is inspired by Zhang and Saligrama (2016b), whose idea is shown in Fig. 1: thanks to the powerful discriminativity of pre-trained CNN, the visual features of test images can be separated into different clusters. We denote the centers of these clusters as *real* centers. We believe that if we have a perfect projection function to project the semantic attributes to the visual space, the projected points (called *synthetic* centers) should align with *real* centers. However, due to the domain shift problem, the projection function learned on the source domain is not perfect so that the *synthetic* centers (*i.e.*VCL centers in Fig. 1) will deviate from *real* centers, and then NN search among these deviated centers to assign labels will cause inferior ZSL performance. Based on the above analysis, besides source domain data, we attempt to take advantage of the existing discriminative structure of target unseen class clusters during the learning of the projection function, *i.e.*the learned projection function should also align the synthetic centers with the real ones in the target domain.

3.1 Visual Center Learning (VCL)

In this section, we first introduce a baseline method which learns the projection function f only with source domain data. Specifically, a CNN feature extractor $\phi(\cdot)$ is used to convert each image x into a d-dimensional feature vector $\phi(x) \in \mathcal{R}^{d\times 1}$. According to the above analysis, each class i of source domain should have a *real* visual center c_i^s, which is defined as the mean of all feature vectors in the corresponding class. For the projection function f, a two-layer embedding network is utilized to transfer source semantic attribute a_i^s to generate corresponding synthetic center $c_i^{syn,s}$:

$$c_i^{syn,s} = f(a_i^s) = \sigma_2(w_2^T \sigma_1(w_1^T a_i^s)) \tag{1}$$

where $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ denote non-linear operation (Leaky ReLU with negative slope of 0.2 by default). w_1 and w_2 are the weights of two fully connected layers to be learned.

Since the correspondence relationship is given in the source domain, we directly adopt the simple mean square loss to minimize the distance between synthetic centers c^{syn} and real centers c in the visual feature space:

$$\mathcal{L}_{MSE} = \frac{1}{S}\sum_{i=1}^{S} \left\| c_i^{syn,s} - c_i^s \right\|_2^2 + \lambda\Psi(w_1, w_2) \tag{2}$$

where $\Psi(\cdot)$ is the $L2$-norm parameter regularizer decreasing the model complexity, we empirically set $\lambda = 0.0005$. Need to note that different from Zhang et al. (2017) which trains with a large number of individual instances of each class i, we choose to utilize a single cluster center c_i^s to represent each object class, and train the model with just several center points. It is based on the observation that the instances of the same category could form compact clusters, and will make our method much more computationally efficient.

When performing ZSL prediction, we first project the semantic attributes of each unseen class i to its corresponding synthetic visual center $c_i^{syn,u}$ using the learned embedding network as in Equation (1). Then for a test image x_k^u, its classification result i^* can be achieved by selecting the nearest synthetic center in the visual space. Formally,

$$i^* = \underset{i}{argmin} \left\| \phi(x_k^u) - c_i^{syn,u} \right\|_2 \tag{3}$$

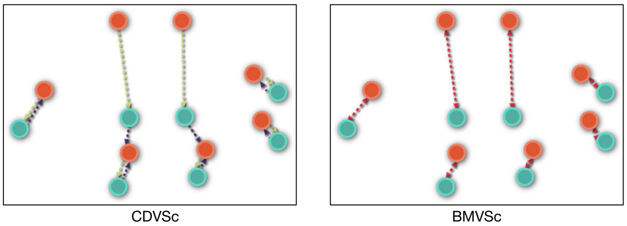

Fig. 2 Illustration of possible many-to-one matching problem in Chamfer-distance based visual structure constraint, which can be avoided in the bipartite matching based visual structure constraint

3.2 Chamfer-Distance-based Visual Structure Constraint(CDVSc)

As discussed earlier, the domain shift problem will cause the target synthetic centers $c^{syn,u}$ deviated from the real centers c^u, thus yields poor performance. Intuitively, if we also require the projected synthetic centers to align with the real ones by using the target domain dataset during the learning process, a better projection function can be learned. However, due to the lack of the label information of the target domain, it is impossible to directly get real centers of unseen classes. Considering the fact that the visual features of unseen classes can be separated into different clusters, we try to utilize some unsupervised clustering algorithms (K-means by default) to get approximated real centers. To valid it, we plot the K-means centers in Fig. 1, which are very close to the real ones.

After obtaining the cluster centers, aligning the structure of cluster centers to that of synthetic centers can be formulated as reducing the distance between the two unordered high-dimensional point sets. Inspired by the work in 3D point clouds (Fan et al. 2017), a symmetric Chamfer-distance constraint is proposed to solve the structure matching problem:

$$\begin{aligned} \mathcal{L}_{CD} = &\sum_{x \in C^{syn,u}} min_{y \in C^{clu,u}} \|x - y\|_2^2 \\ &+ \sum_{y \in C^{clu,u}} min_{x \in C^{syn,u}} \|x - y\|_2^2 \end{aligned} \tag{4}$$

where $C^{clu,u}$ indicates the cluster centers of unseen classes obtained by K-means algorithm. $C^{syn,u}$ represents the synthetic target centers obtained with the learned projection. Combining the above constraint, the final loss function to train the embedding network is defined as:

$$\mathcal{L}_{CDVSc} = \mathcal{L}_{MSE} + \beta \times \mathcal{L}_{CD} \tag{5}$$

3.3 Bipartite-Matching-based Visual Structure Constraint(BMVSc)

CDVSc helps to preserve the structure similarity of two sets, but sometimes many-to-one matching may happen with the Chamfer-distance constraint as shown in Fig. 2. This conflicts with the important prior in ZSL that the obtained matching relation between synthetic and real centers should conform to the strict one-to-one principle. When undesirable many-to-one matching arises, the synthetic centers will be pulled to incorrect real centers and result in inferior performance. To address this issue, we change CDVSc to bipartite matching based visual structure constraint (BMVSc), which aims to find a global minimum distance between the two sets meanwhile to satisfy the strict one-to-one matching principle.

We first consider a graph $G = (V, E)$ with two partitions A and B, where A is the set of all synthetic centers $Csyn, u$ and B contains all cluster centers of target classes. Let $dis_{ij} \in D$ denotes the Euclidean distance between $i \in A$ and $j \in B$, element x_{ij} of the assignment matrix X defines the matching relationship between i and j. To find a one-to-one minimum matching between real and synthetic centers, we could formulate it as a min-weight perfect matching problem, and optimize the problem as follows:

$$\begin{aligned} &\mathcal{L}_{BM} = \min_X \sum_{i,j} dis_{ij} x_{ij}, \\ &s.t. \quad \sum_j x_{ij} = 1, \sum_i x_{ij} = 1, x_{ij} \in \{0, 1\} \end{aligned} \tag{6}$$

In this formulation, the assignment matrix X strictly conforms to the one-to-one principle. To solve this linear programming problem, we employ Kuhn-Munkres algorithm whose time complexity is $O(V^2E)$. Like **CDVSc**, we also combine the MSE loss and this bipartite matching loss.

$$\mathcal{L}_{BMVSc} = \mathcal{L}_{MSE} + \beta \times \mathcal{L}_{BM} \tag{7}$$

3.4 Wasserstein-Distance-based Visual Structure Constraint(WDVSc)

Ideally, if the synthetic and real centers are compact and accurate, the above bipartite matching based distance can achieve a global optimal matching. However, this assumption is not always valid, especially for the approximated cluster centers of target classes, because these centers may contain noises and are not accurate enough. Therefore, instead of using a hard-value (0 or 1) assignment matrix X, a soft-value X whose values represent the joint probability distribution between these two point sets is further considered by using the Wasserstein distance. In the optimal transport theory, Wasserstein distance is demonstrated as a good metric for measuring the distance between two discrete distributions, whose goal is to find the optimal "coupling matrix" X that achieves the minimum matching distance. Its objective formulation is the same as Equation (6), but X represents the soft joint probability values rather than $\{0, 1\}$. In this paper, in order to make this optimization problem convex and solve it more efficiently, we adopt the entropy-regularized optimal transport problem by using the Sinkhorn iterations (Cuturi 2013).

$$\mathcal{L}_{WD} = \min_{X} \sum_{i,j} dis_{ij} x_{ij} - \epsilon H(X)$$
$$H(X) = -\sum_{ij} x_{ij} log x_{ij} \tag{8}$$

where $H(X)$ is the entropy of matrix, ϵ is the regularization coefficient to encourage smoother assignment matrix X. The solution X can be written in form $X = diag\{u\}Kdiag\{v\}$ ($diag\{v\}$ returns a square diagonal matrix with vector v as the main diagonal), and the iterations alternate between updating u and v is:

$$u^{(k+1)} = \frac{a}{Kv^{(k+1)}}, \qquad v^{(k+1)} = \frac{b}{K^T u^{(k+1)}} \tag{9}$$

Here, K is a kernel matrix calculated with D. Since these iterations are solving a regularized version of the original problem, the corresponding Wasserstein distance that results is sometimes called the Sinkhorn distance. Combining this constraint, the final loss function is:

$$\mathcal{L}_{WDVSc} = \mathcal{L}_{MSE} + \beta \times \mathcal{L}_{WD} \tag{10}$$

3.5 A Realistic Setting with Unrelated Test Data in the Wild

Existing transductive ZSL methods always assume that all the images in the test dataset belong to target unseen classes we have already defined. However, in the wild, many unrelated images which do not belong to any defined class may exist. If we directly perform clustering on all these unfiltered images, the approximated real centers will deviate far from the real centers of unseen classes and make the proposed visual structure constraint invalid. This problem also exists in Zhang and Saligrama (2016b). To solve this realistic challenging setting, we propose two new training strategies for our method. In both of these two strategies, we assume the domain shift problem is not that severe, so the initial unseen synthetic centers learned by VCL are roughly correct. This assumption makes sense in that the source and target classes are related in a common attribute space.

Specifically, in the first strategy, we use the baseline VCL to get the initial unseen synthetic centers and then directly filter out unrelated images with one specific feature distance threshold before conducting **CDVSc**, **BMVSc** or **WDVSc**. The details of this training strategy are shown in Algorithm 1. Since the visual structure constraint is only leveraged in one step, this strategy is called the "one-step" training strategy. Since we do not have the ground truth labels of target domain and many noisy samples exist, we use the farthest sample feature distance of the source domain to set the feature distance threshold λ_{dist}.

Algorithm 1 The "one-step" training strategy for the realistic setting in the wild.

Step 1: Only use the source domain data to train the baseline **VCL** to learn an initial projection function f.

Step 2: For each source class in $\mathcal{Y}_s$ in the visual feature space, we find its farthest sample feature point pair and calculate their feature distance $\mathcal{D}_i$. And we set the distance threshold as $\lambda_{dist} = max\{\mathcal{D}_i\}/2$

Step 3: Map each target attribute a_i^u to get its corresponding synthetic visual center $c_i^{syn,u}$, and we denote $C = \{c_1^{syn,u}, ..., c_U^{syn,u}\}$. Then select reliable images x if and only if $\exists c_i \in C, \|x - c_i\|_2^2 \leq \lambda_{dist}$ to construct a new target domain and perform unsupervised clustering on this domain

Step 4: Conduct **CDVSc**, **BMVSc** or **WDVSc**.

By contrast, in the second strategy, rather than using only one specific feature distance threshold, we do it in a progressive way which aims to get more accurate synthetic centers and use larger feature distance threshold in multiple alternative steps. At each step t, we will not only use the latest projection function f_{t-1} to get more accurate synthetic visual centers but also use looser feature distance threshold to include more target domain data. The following experiments will demonstrate that this progressive training strategy is often more robust and can generate better results.

4 Experiments

Implementation Details. We adopt the pretrained ResNet-101 to extract visual features unless specified. All images are resized to 224 × 224 without any data augmentation, and the dimension of extracted features is 2048. The hidden unit numbers of the two FC layers in the embedding network are both 2048. Both visual features and semantic attributes are L2-normalized. Using Adam optimizer, our method is trained for 5000 epochs with a fixed learning rate of 0.0001. The weight β in CDVSc and BMVSc is cross-validated in $[10^{-4}, 10^{-3}]$ and $[10^{-5}, 10^{-4}]$ respectively, while WDVSc directly sets $\beta = 0.001$ because of its very stable performance.

Algorithm 2 The "progressive" training strategy for the realistic setting in the wild. T is the alternative step number.

```
Step 1: Similar as the "one-step" training strategy, we first only use the source domain data to train the baseline VCL to learn an initial projection function $f_0$.

Step 2: Similarly, for each source class in $\mathcal{Y}_s$ in the visual feature space, we find its farthest sample feature point pair and calculate their feature distance $\mathcal{D}_i$. And we set the maximum distance threshold as $\lambda_{dist}^{max} = max\{\mathcal{D}_i\}/2$.

Step 3: We further classify all the target domain samples by finding the closest $c_i^{syn,u}$ as Equation (3). And for each target unseen class, we find the nearest feature sample to the synthetic visual center $c_i^{syn,u}$ and denote the minimum feature distance as $\lambda_{min}^i$.

for $t$ = 1 to T do

    Step t-1: Map each target attribute $a_i^u$ to get its corresponding synthetic visual center $c_i^{syn,u}$ using the projection function $f_{t-1}$, and we denote $C_t = \{c_1^{syn,u}, ..., c_U^{syn,u}\}$.

    Step t-2: For each $c_i^{syn,u}$, we set the distance threshold $\lambda_{dist}^{i,t} = \lambda_{min}^i + (t-1) * (\lambda_{dist}^{max} - \lambda_{min}^i)$. Then select reliable images $x$ if $\|x - c_i^{syn,u}\|_2^2 \leq \lambda_{dist}^{i,t}$ to construct a new target domain.

    Step t-3: Conduct CDVSc, BMVSc or WDVSc to learn an updated projection function $f_t$.
end for
```

Datasets. Extensive experiments are conducted on three widely-used ZSL benchmark datasets, *i.e.*, Animals with Attributes 2 (AwA2) (Xian et al. 2018), Caltech-UCSD Birds 200-2011 (CUB) (Wah et al. 2011) and Scene UNderstanding (SUN) (Patterson et al. 2014). The statistics of these datasets are briefly introduced as below:

- **Animals with Attributes2 (AwA2)** (Xian et al. 2018) contains 37,322 images from 50 animals categories, where 40 of 50 classes are used for training and the rest 10 are used for testing. We adopt the class-level continuous 85-dim attributes as the semantic representations. For fair comparison with previous methods, we also report the results on **AwA1** Lampert et al. (2014) which is an old version of animal datasets of ZSL.
- **Caltech-UCSD Birds 200-2011 (CUB)** (Wah et al. 2011) is a fine-grained bird dataset with 200 species of birds and 11,788 images. Each image is associated with a 312-dim continuous attribute vector. Following Xian et al. (2018), we use the class-level attribute vector and the 150/50 split.
- **SUN-Attribute (SUN)** (Patterson et al. 2014) includes 14,340 images coming from 717 scenes. Each sample is paired with a binary 102-dim semantic vector. We compute class-level continuous attributes as our semantic representations by averaging the image-level attributes for each class. 707/10 (**SUN10**) and 645/72 (**SUN72**) splits are adopted in our experiments.

Data Splits. (1) **Standard Splits (SS)**: The standard seen/unseen class split is first proposed in Lampert et al. (2009) and then widely used in most ZSL works. (2) **Proposed Splits (PS)**: Many recent ZSL methods extract visual features using ImageNet 1K classes pretrained CNN models, and the unseen classes in SS may overlap with these 1K classes, which actually violates the zero-shot setting that the test classes should be *unseen* during ZSL training. Based on this consideration, Xian et al. (2018) introduces the Proposed Splits (PS), in which the overlapped ImageNet classes are removed from the test set of unseen classes. In this paper, we report the results on both the standard splits and the proposed splits for fair comparisons.

Evaluation Metrics. For fair comparison and completeness, we consider two different ZSL settings: 1) **Conventional ZSL**, which assumes all the test instances only belong to target unseen classes. 2) **Generalized ZSL**, where test instances are from both seen and unseen classes, which is a more realistic setting in real applications. For the former setting, we compute the multi-way classification accuracy (MCA) as in previous works.

$$accy = \frac{1}{\|\mathcal{Y}\|}\sum_{i=1}^{\|\mathcal{Y}\|}\frac{\text{\# correct predictions in i}}{\text{\# samples in i}} \tag{11}$$

For the latter one, we define three metrics. 1) $accy_{\mathcal{Y}_s}$ – the accuracy of classifying the data samples from the seen classes to all the classes (both seen and unseen); 2) $accy_{\mathcal{Y}_u}$ – the accuracy of classifying the data samples from the unseen classes to all the classes; 3) H – the harmonic mean of $accy_{\mathcal{Y}_s}$ and $accy_{\mathcal{Y}_u}$:

$$H = \frac{2 * accy_{\mathcal{Y}_u} * accy_{\mathcal{Y}_s}}{accy_{\mathcal{Y}_u} + accy_{\mathcal{Y}_s}} \tag{12}$$

4.1 Qualitative Results

In Fig. 3, we have shown some qualitative results of the proposed **BMVSc** on the AwA2 and CUB datasets. Although the test images of each class have an overall different appearance, the projection function learned by our method can still capture important discriminative semantic information from their visual characteristics, which corresponds to their semantic attributes. For example, the predicted sheep images in AwA2 all share *furry*, *bulbous* and *hooves* attributes. However, we could also observe some misclassified images such as the walrus in row 6 of AwA2. After careful analysis, we find two possible reasons: 1) The discriminative ability of the

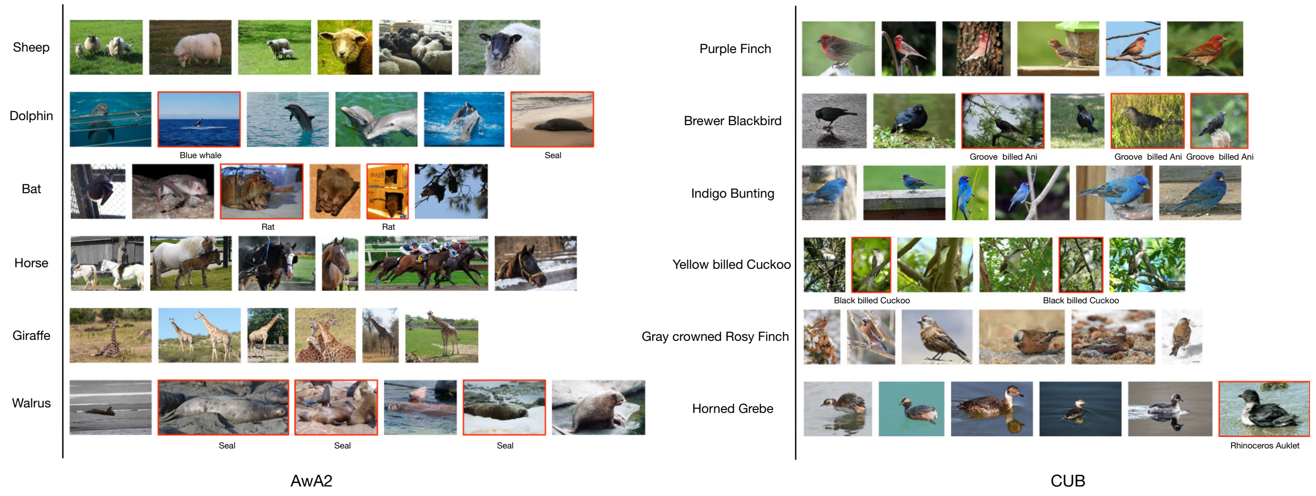

Fig. 3 Qualitative results of BMVSc on 6 categories of AwA2 and CUB datasets. We list the top-6 images classified to each category. The misclassified images are marked with red bounding boxes and the right name of category is below the corresponding image

pretrained CNN is not enough to separate the visual appearances between highly similar categories. In fact, the visual appearance of seal and walrus are so close that even people could not distinguish them by rule and line without expert knowledge. This problem can be solved only by more powerful visual features. 2) Some attribute annotations are not accurate enough. For example, the seal category possesses *spots* of semantic descriptions, but walrus does not, but both these two categories own this attribute in the semantic annotation. Such incorrect supervision information will mislead the learning of the projection function.

4.2 Conventional ZSL Results

To show the effectiveness of the proposed visual structure constraint, we first compare our method with existing state-of-the-art methods in the conventional setting. Table 1 is the comparison results under standard splits (**SS**), where we also re-implement our method using 4096-dimensional VGG features to guarantee fairness. Obviously, with the three different types of visual structure constraint, our method can obtain substantial performance gains consistently on all the datasets and outperforms previous state-of-the-art methods. The only exception is that VZSL (Wang et al. 2018) is slightly better than our method on the AwA1 dataset when using VGG features.

Specially, comparing with SP-ZSR (Zhang and Saligrama 2016b) which shares the similar spirit with our method, we could find that their performance sometimes is even worse than inductive methods such as SynC (Changpinyo et al. 2016), SCoRe (Morgado and Vasconcelos 2017) or VCL. The possible underlying reason is that, when utilizing the structure information only in test time, the final performance gain highly depends on the quality of the projection function. When the projection function is not good enough, the initial synthetic centers will deviate far from the real centers and result in bad matching results with unsupervised cluster centers, thus causing even worse results. By contrast, in our method, this visual structure constraint is incorporated into the learning of projection function in the training stage, which can help to learn a better projection function and bring performance gain consistently. Another bonus is that, during runtime, we can directly do recognition in real-time online-mode rather than the batch-mode optimization in SP-ZSR (Zhang and Saligrama 2016b), which is more friendly in real applications.

The results on proposed splits of AwA2, CUB and SUN72 are reported in Table 2 with ResNet-101 features. It can be seen that almost all methods suffer from performance degradation under this setting. However, our proposed method could still maintain the highest accuracy. Specifically, the improvements obtained by our method range from 0.8% to 25.8%, which indicate that visual structure constraint is effective in solving the domain shift problem.

4.3 Large-Scale Conventional ZSL Results

To further evaluate the effectiveness of the real large-scale scenarios, we follow the procedure of Frome et al. (2013) and evaluate our method on the widely used subset of ImageNet. It corresponds to 1,549 unseen classes that are within two hops of the 1,000 seen classes according to the existing hierarchical structure of ImageNet. There exists certain classes without specific semantic representations. Instead of dropping them directly like Changpinyo et al. (2018), we average each word embedding to construct corresponding semantic attributes. For the evaluation metric, we employ Hit@K as Frome et al. (2013) which is defined as the percentage of test

Table 1 Quantitative comparisons of MCA (%) under standard splits (SS) in conventional ZSL setting. **I**: Inductive, **T**: Transductive, **O**: Our method, Bold: Best, Italic: Second best, V: VGG, R: ResNet, G: GoogLeNet. It can be seen that the proposed three types of visual structure constraints can bring substantial performance gain and outperform existing state-of-the-art methods

	Method	Features	AwA1	AwA2	CUB	SUN72	SUN10
I	CONSE Norouzi et al. (2014)	R	63.6	67.9	36.7	44.2	–
	SSE Zhang and Saligrama (2015)	V	76.3	–	30.4	–	82.5
	JLSE Zhang and Saligrama (2016a)	V	80.5	–	42.1	–	83.8
	SynC Changpinyo et al. (2016)	R	72.2	71.2	54.1	59.1	–
	SAE Kodirov et al. (2017)	R	80.6	80.7	33.4	42.4	–
	SCoRe Morgado and Vasconcelos (2017)	V	82.8	–	59.5	–	–
	f-CLSWGAN Xian et al. (2018)	R	69.9	–	61.5	62.1	–
T	SP-ZSR Zhang and Saligrama (2016b)	V	92.0	–	53.2	–	86.0
	DSRL Ye and Guo (2017)	V	87.2	–	57.1	–	85.4
	DMaP Li et al. (2017)	V+G+R	90.5	–	67.7	–	–
	VZSL Wang et al. (2018)	V	94.8	–	66.5	–	87.8
	QFSL Song et al. (2018)	V	–	84.1	61.2	–	–
O	**VCL**	V	81.7	82.6	58.2	58.8	87.2
	CDVSc	V	89.6	93.3	69.9	59.7	90.6
	BMVSc	V	92.7	94.0	70.8	61.3	89.7
	WDVSc	V	92.9	94.2	71.0	62.3	91.2
	VCL	R	82.0	82.5	60.1	63.8	89.6
	CDVSc	R	94.3	93.9	**74.2**	64.5	90.5
	BMVSc	R	*95.9*	**96.8**	*73.6*	*66.2*	*91.7*
	WDVSc	R	**96.2**	*96.7*	**74.2**	**67.8**	**92.2**

Table 2 Quantitative comparisons under the proposed splits (PS) in conventional ZSL setting. It shows that, though most methods suffer from performance degradation under this setting, our method can still achieve the best performance and beat other methods by a large margin

Method	AwA2	CUB	SUN72	Ave.
CONSE Norouzi et al. (2014)	44.5	34.3	38.8	39.2
DeViSE Frome et al. (2013)	59.7	52.0	56.5	56.0
SJE Akata et al. (2015)	61.9	53.9	53.7	56.5
SynC Changpinyo et al. (2016)	46.6	55.6	56.3	52.8
SAE Kodirov et al. (2017)	54.1	33.3	40.3	42.5
SCoRe Morgado and Vasconcelos (2017)	69.5	61.0	51.7	60.7
LDF Li et al. (2018a)	–	69.2	–	–
cycle-CLSWGAN Felix et al. (2018)	66.3	58.4	60.0	61.6
PSR-ZSL Annadani and Biswas (2018)	63.8	56.0	61.4	60.4
LisGAN Li et al. (2019)	70.6	58.8	61.7	63.7
DCN Liu et al. (2018b)	–	56.2	61.8	–
LsrGAN Li et al. (2019)	66.4	60.3	*62.5*	63.1
VCL	61.5	59.6	59.4	60.1
CDVSc	78.2	*71.7*	61.2	70.3
BMVSc	*81.7*	71.0	62.2	*71.6*
WDVSc	**87.3**	**73.4**	**63.4**	**74.7**

Table 3 Quantitative results on the large-scale ImageNet dataset in conventional ZSL setting. Even under this very challenging setting, our method still beats the baseline method VCL and previous state-of-the-art methods by more than 5%

Test data	Method	Hit@K (%)				
		1	2	5	10	20
ILSVRC2010	**DIPL** Zhao et al. (2018)	–	–	31.7	–	–
2-hop	CONSE Norouzi et al. (2014)	8.3	12.9	21.8	30.9	41.7
	SYNC Changpinyo et al. (2016)	10.5	16.7	28.6	40.1	52.0
	EXEM Changpinyo et al. (2017)	12.5	19.5	32.3	43.7	55.2
	f-CLSWGAN Xian et al. (2018)	10∼12	–	–	–	–
	f-VAEGAN-D2 Xian et al. (2019)	13∼14	–	–	–	–
2-hop	**VCL**	12.3	19.3	31.3	40.3	48.7
	WDVSc	**17.6**	**26.7**	**38.8**	**47.5**	**57.9**

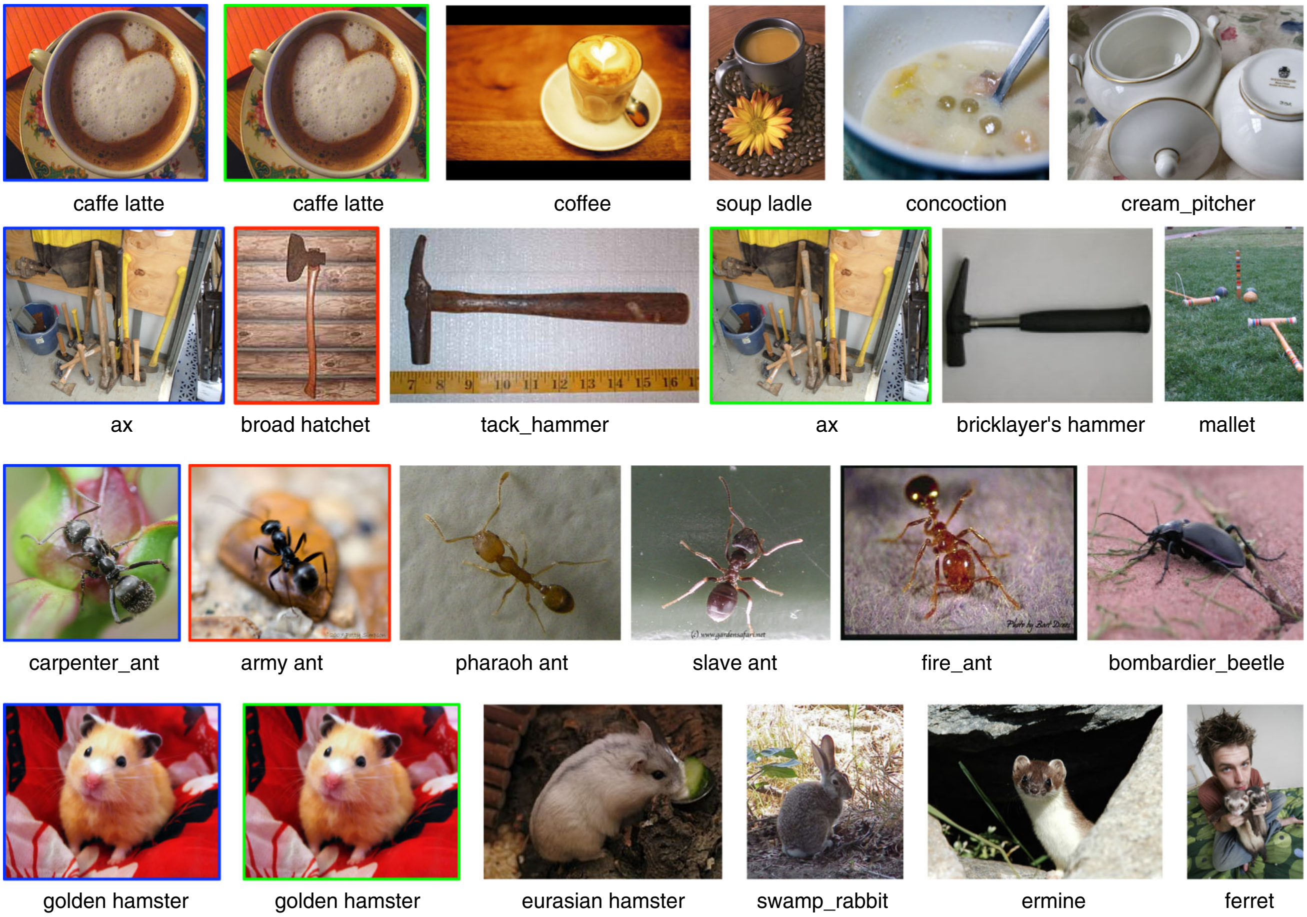

Fig. 4 Some classification results on the large-scale imagenet dataset. The leftmost column of the images are the input test images, while the second to six columns of images are representative sample images of top-5 classified categories. It shows that though the top-1 classification results of some samples are not correct, they are all predicted into visually similar categories

images whose true labels belong to top K predictions of models. As shown in Table 3, our WDVSc outperforms baseline VCL and previous state-of-the-art methods by more than 5%. Note that the baseline method DIPL (Zhao et al. 2018) only reports its results on the test data ILSVRC 2010, which is a subset of the default 2-hop test data, so it is relatively easier.

In Fig. 4, we further provide four classification results. Though the number of unseen classes is very large in this dataset, our method can still work quite well and correctly categorize the input test images into corresponding categories. Moreover, for some samples whose top-1 classification results are not correct, they are all categorized into visually similar categories. It is consistent with the increased

Table 4 Quantitative comparisons under the generalized ZSL setting. Obviously, most methods including our baseline VCL cannot maintain the same level accuracy for both seen and unseen classes. The much better results on the source seen classes should be because of the serious domain shift problem, which can be significantly alleviated by the proposed visual structure constraints

Method	AwA2			CUB			SUN72		
	$accy_u$	$accy_s$	H	$accy_u$	$accy_s$	H	$accy_u$	$accy_s$	H
CONSE Norouzi et al. (2014)	0.5	**90.6**	1.0	1.6	72.2	3.1	6.8	39.9	11.6
SSE Zhang and Saligrama (2015)	8.1	82.5	14.8	8.5	46.9	14.4	2.1	36.4	4.0
DeViSE Frome et al. (2013)	17.1	74.7	27.8	23.8	53.0	32.8	16.9	27.4	20.9
SJE Akata et al. (2015)	8.0	73.9	14.4	23.5	59.2	33.6	14.7	30.5	19.8
ESZSL Romera-Paredes and Torr (2015)	5.9	77.8	11.0	12.6	63.8	21.0	11.0	27.9	15.8
SynC Changpinyo et al. (2016)	10.0	90.5	18.0	11.5	70.9	19.8	7.9	43.3	13.4
ALE Akata et al. (2016)	14.0	81.8	23.9	23.7	62.8	34.4	21.8	33.1	26.3
PSR-ZSL Annadani and Biswas (2018)	20.7	73.8	32.3	24.6	54.3	33.9	20.8	37.2	26.7
cycle-CLSWGAN Felix et al. (2018)	56.9	64.0	60.2	45.7	61.0	52.3	**49.4**	33.6	40.0
LisGAN Li et al. (2019)	52.6	76.3	62.3	46.5	57.9	51.6	42.9	37.8	40.2
f-VAEGAN-D2 Xian et al. (2019)	57.6	70.6	63.5	**48.4**	60.1	*53.6*	45.1	38.0	**41.3**
LsrGAN Vyas et al. (2020)	54.6	74.6	63.0	48.1	59.1	53.0	44.8	37.7	40.9
VCL	21.4	89.6	34.6	15.6	**86.3**	26.5	10.4	**63.4**	17.9
CDVSc	66.9	88.1	76.0	37.0	84.6	51.4	27.8	63.2	38.6
BMVSc	71.9	88.2	*79.2*	33.1	86.1	47.9	29.9	62.9	40.6
WDVSc	**76.4**	88.1	**81.8**	43.3	85.4	**57.5**	30.5	63.1	*41.1*

performance number with a larger K. In fact, it is even difficult for human to tell the difference among these visually similar categories.

4.4 Generalized ZSL Results

In Table 4, we compare our method with eight different generalized ZSL methods. It can be seen that, although almost all the methods including our baseline VCL cannot maintain the same level accuracy for both seen ($accy_s$) and unseen classes ($accy_u$), our method with visual structure constraints significantly outperforms other methods by a large margin on these datasets. More specifically, take CONSE (Norouzi et al. 2014) as an example, due to the domain shift problem, it can achieve the best results on the source seen classes but totally fails on the target unseen classes. By contrast, since the proposed two structure constraints can help to align the structure of synthetic centers to that of real unseen centers, our method can achieve acceptable ZSL performance on target unseen classes. This generalized ZSL setting demonstrates that the proposed visual structure constraints can help alleviate the aforementioned domain shift problem again.

4.5 Results of the Realistic Setting in the Wild

To imitate the setting in the wild where many unrelated images may exist in the test dataset, we prepare two types of datasets, coarse-grained and fine-grained respectively. Specifically, for the coarse-grained dataset, we mix the AwA2 test dataset with extra 8K unrelated images from the aPY dataset. These unrelated images do not belong to the classes of either AwA2 or ImageNet-1K. For the fine-grained dataset, we hold out 10 unseen classes of CUB dataset as noise samples to confuse the target domain. From Table 5, we have the following analyses. 1)It could be seen that without filtering out the unrelated images, the performance of our method with CDVSc, BMVSc, and WDVSc all degrades when compared to the baseline VCL on the coarse-grained dataset, which means that the alignment of wrong visual structures is counterproductive to the learning of projection function. 2) On the fine-grained dataset, directly constraining the noisy target domain with the visual center does not hurt that much. The underlying reason may be that the visual feature of the fine-grained image sometimes is indistinguishable as shown in Fig. 9. 3) With the new training strategies ($S+*$, $P+*$), the proposed visual structure constraints can work very well and bring substantial performance gain consistently on both coarse-grained and fine-grained datasets. 4) The **progressive training strategy** $P+*$ is better than the **one-step training strategy** $S+*$, which demonstrates the benefits from gradually improving the projection function with more domain target samples and looser distance constraints (Table 6).

Besides the final performance, we further analyze the influence of distance threshold on these two strategies for the standard split (top) and the proposed split (bottom) in Fig. 5 on AwA2 dataset. Though they both have different MCA results for different threshold values, we find the progressive strategy is also overall more robust and better than

Table 5 Results (%) in the more realistic setting. With the new proposed one-stage ($S + *$) and progressive ($P + *$) training strategy, the proposed method can still work well and bring performance gain. In particular, we find WDVSc is extremely robust in this challenging setting thanks to its inherent soft matching scheme and the progressive training strategy is better than the one-stage one as well. The experiments are conducted on both coarse-grained dataset AwA and fine-grained dataset CUB

Method	VCL	CDVSc	BMVSc	WDVSc	S+CDVSc	S+BMVSc	S+WDVSc	P+CDVSc	P+BMVSc	P+WDVSc
AwA SS+noise	82.5	79.7	78.3	81.3	89.3	86.9	92.4	90.0	88.5	**92.5**
AwA PS+noise	61.5	57.4	58.9	60.8	65.3	66.7	78.3	73.7	75.4	**79.3**
CUB SS+noise	63.0	62.2	63.9	63.4	65.2	65.3	67.7	65.9	66.9	**68.1**
CUB PS+noise	68.2	70.8	69.3	68.1	72.3	70.6	72.9	**74.2**	70.9	73.6

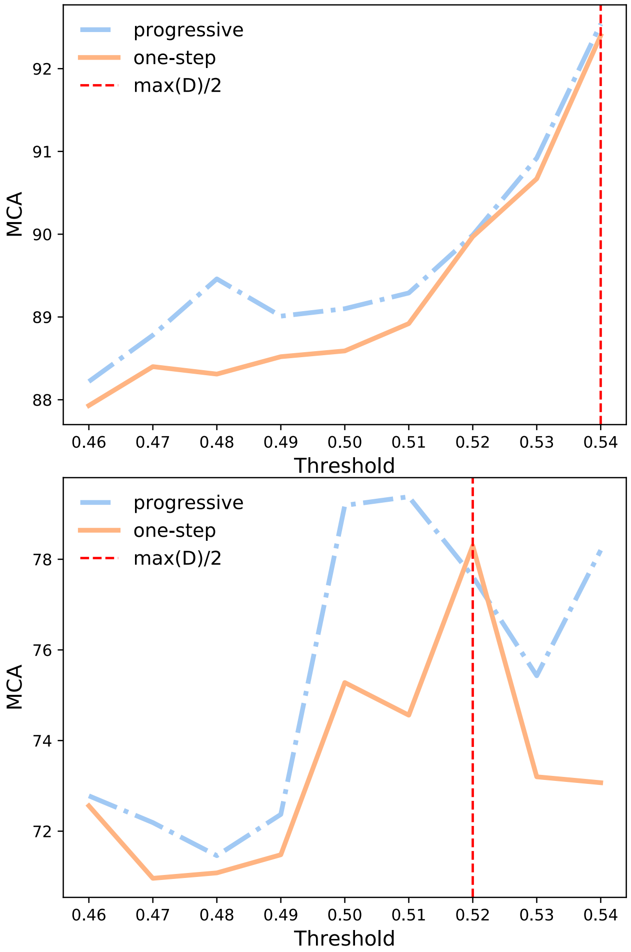

Fig. 5 The comparison results between the one-step and progressive training strategy in the realistic setting for standard split (SS, top) and proposed split (PS, top) respectively. The x-axis threshold denotes the feature distance threshold and the max feature distance threshold in these two settings respectively. It shows that the progressive training strategy is overall more robust and better than the one-step one for different thresholds

the one-step training strategy. Need to note that the threshold value here means the feature distance threshold λ_{dist} in Algorithm 1 and the maximum distance threshold λ_{dist}^{max} in Algorithm 2 respectively.

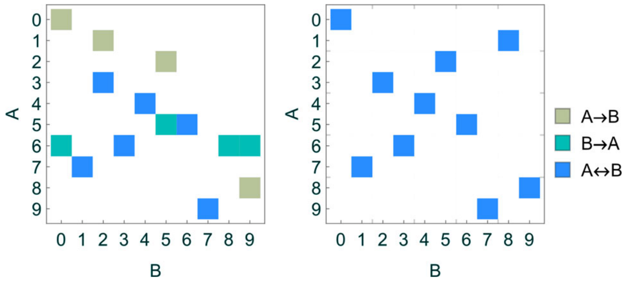

Fig. 6 Matching matrixs between the projected semantic centers and visual cluster centers of **CDVSc** (left) and **BMVSc** (right) on the AwA2 dataset. **BMVSc** can guarantee strict one-one matching while **CDVSc** may have many-to-one matching

4.6 More Analysis

Possible many-to-one problem in CDVSc. To verify that there may exist many-to-one matching problem during the training of **CDVSc**, we randomly select the output of embedding networks of one epoch and visualize the matching results on the AwA2 dataset in Fig. 6. It can be seen that one projected semantic center can be matched by multiple visual cluster centers, and vice versa. By contrast, **BMVSc** can guarantee strict one-one matching, which may be the reason of better results of this dataset.

Progressive improvement of center matching in BMVSc. The final ZSL performance depends on the alignment of the projected semantic centers and real visual centers. In our method we use K-means cluster centers to approximate the real centers and minimize their matching distance. So one natural question would be ”whether we can achieve this final objective by training with cluster centers from K-means?”. To answer this question, we calculate the the number of right matching point and distances between the projected semantic centers and real visual centers respectively during the training of **BMVSc** . We plot these two metrics of the SUN dataset in Fig. 7. Obviously, **BMVSc** can definitely improve the matching of the projected semantic centers and real visual centers by only using the cluster centers from K-means during the whole training processing.

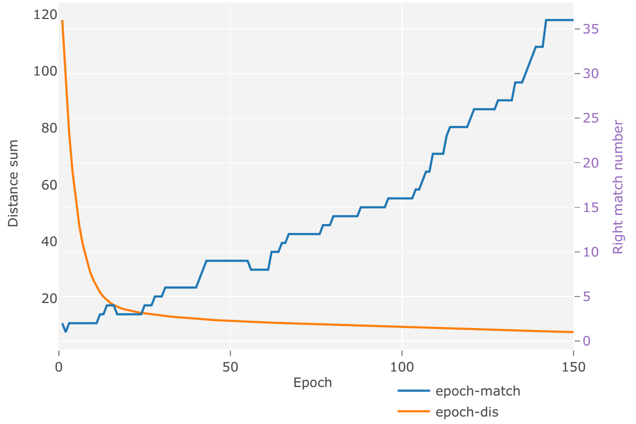

Fig. 7 The right matching number ("epoch-match") and distance ("epoch-dis") between the projected semantic centers and real visual centers during the training of BMVSc on the SUN dataset, which demonstrate the BMVSc can improve the matching relation by only using the approximated K-means centers

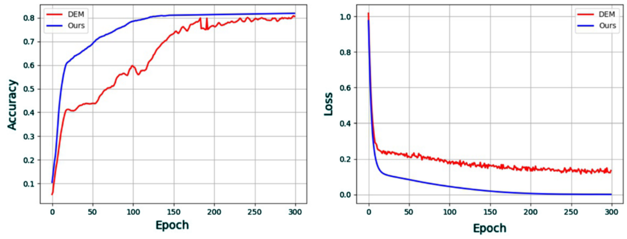

Fig. 8 The comparison of convergence curve between instance-based method (DEM) and our center-based method (VCL). It shows that center-based objective converges faster than instance-based objective

Center-based objective vs Instance-based objective. Compared to previous instance-based optimization objective, our center-based optimization objective is much more computationally efficient. To verify this point, we re-implement the work **DEM** (Zhang et al. 2017) and adopt the same network structure, parameter setting and optimization algorithm with our VCL method on the AwA2 dataset. Then we plot the change of loss and accuracy with epoch increasing in Fig. 8 respectively. It shows that our center-based optimization objective converges faster than previous instance-based optimization objective and can even achieve slightly better final results.

Why are slightly worse results obtained by BMVSc than CDVSc on the CUB dataset? In our paper, three different types of visual structure constraint are proposed to alleviate the domain shift problem in ZSL. BMVSc can solve the possible many-to-one matching problem in CDVSc and satisfy the strict one-to-one principle, which potentially helps to achieve better results, such as the gain can be observed on the AwA2 and SUN datasets. However, on the CUB dataset, the performance of BMVSc is slightly worse than CDVSc. Although this difference is quite subtle when it is compared to the absolute gain coming from the visual structure constraint, we still want to find the possible underlying reason.

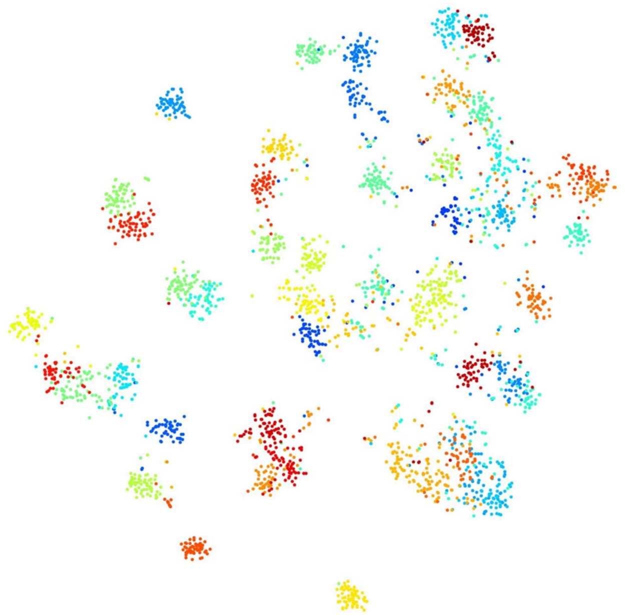

Fig. 9 Feature distribution of the CUB dataset, which shows the feature distribution is not that separable for all categories

To answer this question, we first plot the feature distribution of all the categories of the CUB dataset in Fig. 9 with TSNE. We could find that the feature distribution of some categories is too close to be distinguished because the feature pretrained on ImageNet is not representative enough for this CUB dataset. This somehow violates our assumption that the separated clusters of unseen classes obtained from pre-trained CNN models are already discriminative, and thus lead to this degradation phenomenon. To verify it, we check the matching matrix obtained by our methods and find that there indeed exists wrong matches due to very closed real centers. Specifically, consider synthetic center X of *yellow billed cuckoo*, and two similar real centers Y and Z of *mangrove cuckoo* and *yellow billed cuckoo* which are shown in Fig. 10. $X - Z$ is the right matching, and $X - Y$ is the wrong matching. In BMVSc, if the wrong matching happens, X will be pulled closer to inaccurate center Y (loss term: $\|X - Y\|_2^2$). By contrast, the contribution of CDVSc to the final loss is $\frac{\|X-Y\|_2^2+\|X-Z\|_2^2}{2}$, which will also force X to approach Z and alleviate the wrong matching problem to some certain degree.

Importance of unsupervised cluster centers and semantic attributes. In our method, to recognize the target domain images, two different types of knowledge are leveraged: unsupervised cluster centers of target domain and semantic attributes. To study the importance of these two components, we design a simple voting algorithm to calculate the upper bound of unsupervised clustering algorithms for ZSL recog-

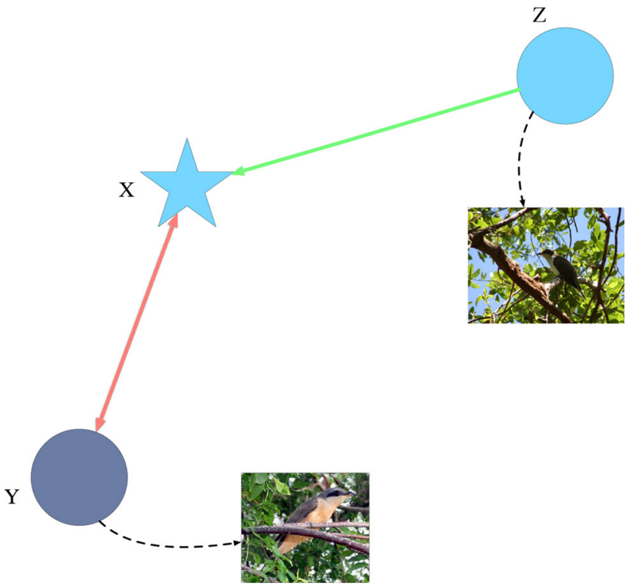

Fig. 10 Matching relations between synthetic center and two similar real centers. Red line denotes BMVSc matching, and green line and red line denote CDVSc matching

Table 7 Analysis to demonstrate the importance of unsupervised cluster centers and semantic attributes. By combining these two types of information during training, our **CDVSc**, **BMVSc** and **WDVSc** achieve better results than the upper bound of **K-Means** and **VCL**

	AwA2	CUB	SUN
K-Means	75.0	67.4	57.6
VCL	61.5	59.6	59.4
CDVSc	78.2	71.7	61.2
BMVSc	81.7	71.0	62.2
WDVSc	87.3	73.4	63.4

nition. Specifically, we assume the ground truth label for each unseen instance is accessible. Then for each cluster center obtained by K-means, we predict its category through a voting process, *i.e.*its category is the one which most images in this cluster belong to. Finally, the classification results for test instances are directly set to the label of the corresponding cluster. In this way, because we have already used the ground truth information, it can be viewed as the upper bound of K-means clustering algorithm. As shown in Table 7, its performance is even better than our baseline **VCL**. which demonstrates that the information of unsupervised clustering is very useful. By combining the semantic attributes and this unsupervised cluster information during the learning process, our method **CDVSc**, **BMVSc** and **WDVSc** are all better than the upper bound of **K-Means** and **VCL**.

Stability of unsupervised cluster centers. Since the proposed the visual structure constraint depends on the unsupervised clustering which has a certain degree of randomness, one may ask "if the final ZSL performance is stable enough"? To analyze this point in detail, we run the whole pipeline for 5 times on both AwA2 and CUB dataset using WDVSc. To eliminate the influence of other factors, we initialize the projection function f with the same parameters, only keeping the randomness of the clustering methods. Based on the experimental results of Table 8, we could find that the performance variance of our method on each dataset is very minor. By using more advanced clustering methods like Chang et al. (2017), we believe the stability and performance could be further improved.

Instance based WDVSc. Though our method uses the default center-based objective function, we find the proposed wassertein-distance-based visual structure constraint can also support instance-based objective function well. This is because the wassertein distance can be used to measure the distance between two discrete distributions with unequal sample numbers, in which case the "coupling matrix" X is not a square matrix anymore. Specifically, we do not use unsupervised clustering algorithms to generate approximated real centers B in advance but instead directly use the visual feature of each instance to find their individual optimal matching. To verify its effectiveness, we conduct the controlled experiment on three different datasets and evaluation settings (standard split "SS" and proposed split "PS"). As shown in Table 9, without unsupervised clustering, the instance-based WDVSc is a little worse than the default center-based WDVSc but still significantly better than the baseline VCL. It further demonstrates the strong generalization ability of the proposed visual structure constraint.

Besides dropping the cluster procedure, instance-based WDVSc also has other advantages compared with center-based WDVSc. Sometimes there may exist some extreme cases where the number of samples in some classes is very small and thus the unsupervised clustering results will be insensible and inaccurate to some extent. In these cases, it is more suitable to use the instance-based WDVSc through

Table 6 Generality to the word vector based semantic space on the AwA1 dataset. Though the word vector contains less effective information than semantic attributes, our visual structure constraint can work very well and bring significant performance gain

Method	DeViSE Frome et al. (2013)	ZSCNN Lei et al. (2015)	SS-Voc Fu and Sigal (2016)	DEM Zhang et al. (2017)	VCL	CDVSc	BMVSc	WDVSc
MCA (%)	50.4	58.7	68.9	78.8	72.3	79.4	83.9	**90.8**

Table 8 We run the whole pipeline multiple times and record the MCA (%) performance on AwA2 and CUB dataset. It demonstrates the stability of used cluster methods. Meanwhile, the averaged results are slightly higher than the previous report results

Random Test	1	2	3	4	5	**Report**	**Mean**	**Std**
AwA2	96.62	96.81	96.72	96.58	96.83	96.7	96.712	± 0.09948
CUB	73.34	73.59	73.76	73.68	73.78	73.4	73.63	± 0.15975

Table 9 Comparison results (%) among the baseline VCL, the default center based WDVSc ("WDVSc") and instance based WDVSc ("WDVSc-instance") on different datasets and conventional settings. It shows that the instance based WDVSc is worse than the default center based WDVSc but still much better than the VCL baseline, which demonstrates the strong generalization ability of the proposed visual structure constraint

Method	AwA2 (SS)	AwA2(PS)	CUB (SS)	CUB (PS)	SUN72 (SS)	SUN72 (PS)
VCL	82.5	61.5	60.1	59.6	63.8	59.4
WDVSc	96.7	87.3	74.2	73.4	67.8	63.4
WDVSc-instance	95.0	82.1	70.3	71.3	65.6	61.4

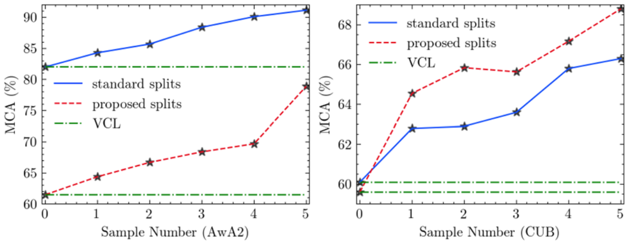

Fig. 11 The comparison results of MCA (%) on the AwA2 and CUB dataset by using instance based WDVSc with different sample number of each target class. It shows that instance based WDVSc can work quite well for this extremely challenging case. And better results can be achieved with more samples

directly measuring the distance of two discrete feature distributions with unequal quantity. To simulate this setting, we test both AwA2 and CUB dataset and only keep 1 to 5 images for each class respectively. It can be seen from Fig. 11 that our method can achieve substantial improvements over the VCL baseline even in this extremely challenging case by using the instance based WDVSc. And with more target images for each class, the final performance increases consistently.

Generality to word vectors based semantic space. Compared to some previous methods which are only applicable to one specific semantic space, we further demonstrate that the proposed visual structure constraint can also be applied to word vector-based semantic space. Specifically, to obtain the word representations for the embedding networks inputs, we use the GloVe text model (Pennington et al. 2014) trained on the Wikipedia dataset leading to 300-d vectors. For the classes containing multiple words, we match all the words in the trained model and average their word embeddings as the corresponding category embedding. As shown in Table 6, though the contained effective information of word vectors is less than that of semantic attributes, the proposed visual structure constraint can still bring substantial performance gain and outperform previous methods. Note that DEM (Zhang et al. 2017) utilized 1000-d word vectors provided by Fu et al. (2014), Fu et al. (2015) to represent a category.

Robustness for imperfect separability of visual features for unseen classes. Though our motivation is inspired by the great separable ability of visual features for unseen classes on the AwA2 dataset, we find the proposed visual structure constraint is very robust and does not rely on it seriously. For example, on the CUB dataset, the feature distribution in Fig. 9 is not totally separable, but the proposed visual structure constraint still brings significant performance gain as shown in the above Tables. Because even though there are some incorrect clusters, as long as most of them are correct clusters, the proposed visual structure constraint will be beneficial.

On the other hand, though the unseen class number K is often pre-defined, we find the proposed visual constraint can improve the performance even when it is unknown. In Fig. 12, we report the performance for different $K(K \leq$ unseen class number) on coarse-grained AwA2 dataset and fine-grained SUN dataset. Specifically, we first perform K-means both in the semantic space and visual space simultaneously, then use WDVSc to align these two synthetic sets. Obviously, the proposed visual structure constraint can bring performance gain consistently. In other words, as long as the visual features can form some superclasses (not fine-level, which is satisfied by most datasets), the proposed visual structure constraint is always effective. We further conduct similar experiments by using WDVSc on the large-scale ImageNet dataset. As shown in Fig. 13, the recognition accuracy by using the visual structure constraint also improves the baseline VCL consistently while increasing the number of visual centers. We believe the proposed visual structure constraint could capture a more fine-grained structure of visual space with an increase of K, thus achieving better results.

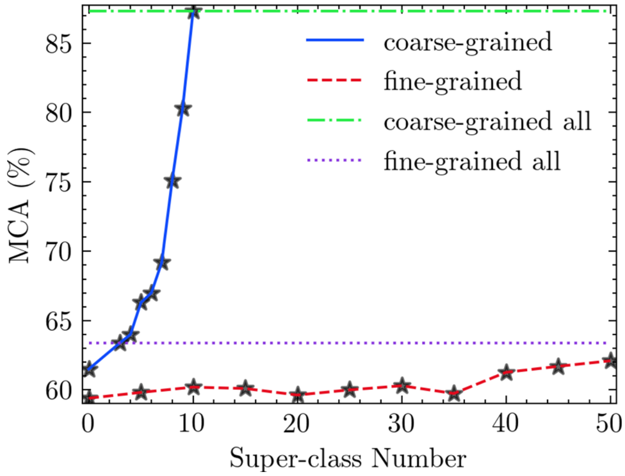

Fig. 12 MCA results (%) of different cluster number K (super-class) on coarse-grained dataset (AwA2) and fine-grained dataset (SUN). Note that the total unseen class number of AwA2 and SUN are 10 and 72 respectively. coarse-grained all and fine-grained all denote that directly use the maximum cluster number

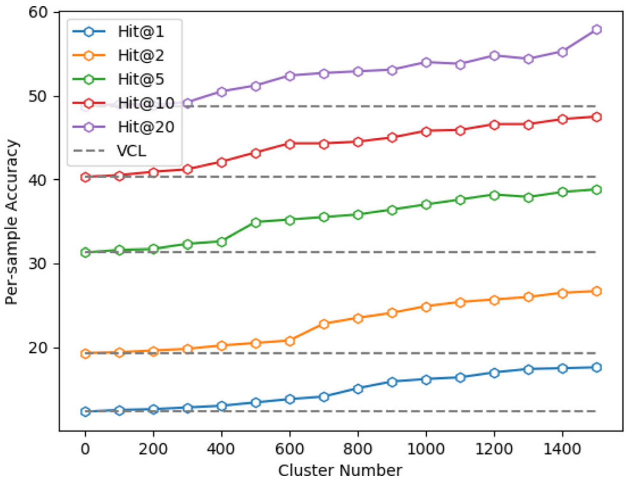

Fig. 13 ImageNet results with different unknown K. On the one hand, it demonstrates the effectiveness of our method even without knowing K. On the other hand, the proposed visual structure constraint could capture more fine-grained structure information with the increase of K and achieves better results. Note that the unseen class number of this setting is 1549

5 Conclusion

Domain shift is a key problem for ZSL in the wild. To alleviate it, three different types of visual structure constraint are proposed for transductive ZSL in this paper. Moreover, we introduce a new transductive ZSL configuration for real scenarios and design two new training strategies to make our method work for it. Experiments demonstrate that our method brings substantial performance gain consistently on different benchmark datasets, outperforms previous state-of-the-art methods by a large margin and generalizes well for data in the wild, including large-scale data and some extreme cases.

Acknowledgements The work described in this paper was fully supported by an ECS grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 21209119).

References

Akata, Z., Perronnin, F., Harchaoui, Z., & Schmid, C. (2013). Label-embedding for attribute-based classification. *In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 819-826)*.

Akata, Z., Perronnin, F., Harchaoui, Z., & Schmid, C. (2016). Label-embedding for image classification. *TPAMI*.

Akata, Z., Reed, S., Walter, D., Lee, H., & Schiele, B. (2015). Evaluation of output embeddings for fine-grained image classification. *In: CVPR*.

Annadani, Y., & Biswas, S. (2018). Preserving semantic relations for zero-shot learning.*In: CVPR*.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein gan. arXiv preprint arXiv:1701.07875.

Chang, J., Wang, L., Meng, G., Xiang, S., & Pan, C. (2017). Deep adaptive image clustering. *In: Proceedings of the IEEE international conference on computer vision, pp. 5879–5887*.

Changpinyo, S., Chao, W.L., Gong, B., & Sha, F. (2016). Synthesized classifiers for zero-shot learning.*In: CVPR*.

Changpinyo, S., Chao, W.L., Gong, B., & Sha, F. (2018). Classifier and exemplar synthesis for zero-shot learning. arXiv preprint arXiv:1812.06423.

Changpinyo, S., Chao, W.L., & Sha, F. (2017). Predicting visual exemplars of unseen classes for zero-shot learning. *In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3476–3485*.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *In: Advances in neural information processing systems, pp. 2292–2300*.

Elhoseiny, M., Saleh, B., & Elgammal, A. (2013). Write a classifier: Zero-shot learning using purely textual descriptions. *In: ICCV*.

Fan, H., Su, H., & Guibas, L. (2017). A point set generation network for 3d object reconstruction from a single image. *In: CVPR*.

Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. (2009). Describing objects by their attributes. *In: CVPR*.

Felix, R., Kumar, V.B., Reid, I., & Carneiro, G. (2018). Multi-modal cycle-consistent generalized zero-shot learning. *In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 21–37*.

Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., & Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. *In: NIPS*.

Fu, Y., Hospedales, T.M., Xiang, T., Fu, Z., & Gong, S. (2014). Transductive multi-view embedding for zero-shot recognition and annotation. *In: ECCV*.

Fu, Y., Hospedales, T.M., Xiang, T., & Gong, S. (2015). Transductive multi-view zero-shot learning. *TPAMI*.

Fu, Y., & Sigal, L. (2016). Semi-supervised vocabulary-informed learning. *In: CVPR*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *In: Advances in neural information processing systems, pp. 2672–2680*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In: CVPR*.

Huang, H., Wang, C., Yu, P.S., & Wang, C.D. (2019). Generative dual adversarial network for generalized zero-shot learning. *In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 801–810*.

Kingma, D.P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kodirov, E., Xiang, T., Fu, Z., & Gong, S. (2015). Unsupervised domain adaptation for zero-shot learning.*In: ICCV*.

Kodirov, E., Xiang, T., & Gong, S. (2017). Semantic autoencoder for zero-shot learning. *In: CVPR*.

Lampert, C.H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. *In: CVPR*.

Lampert, C.H., Nickisch, H., & Harmeling, S. (2014). Attribute-based classification for zero-shot visual object categorization. *TPAMI*.

Lei Ba, J., Swersky, K., & Fidler, S., et al. (2015). Predicting deep zero-shot convolutional neural networks using textual descriptions. *In: ICCV*.

Li, J., Jing, M., Lu, K., Ding, Z., Zhu, L., & Huang, Z. (2019). Leveraging the invariant side of generative zero-shot learning. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7402–7411*.

Li, Y., Wang, D., Hu, H., Lin, Y., & Zhuang, Y. (2017). Zero-shot recognition using dual visual-semantic mapping paths. *In: CVPR*.

Li, Y., Zhang, J., Zhang, J., & Huang, K. (2018). Discriminative learning of latent features for zero-shot recognition. *In: CVPR*.

Liu, S., Long, M., Wang, J., & Jordan, M.I. (2018). Generalized zero-shot learning with deep calibration network.*In: Advances in Neural Information Processing Systems, pp. 2005–2015*.

Long, Y., Liu, L., Shen, F., Shao, L., & Li, X. (2017). Zero-shot learning using synthesised unseen visual data with diffusion regularisation. *IEEE transactions on pattern analysis and machine intelligence*, *40*(10), 2498–2512.

Lu, Y. (2016). Unsupervised learning on neural network outputs: with application in zero-shot learning. *IJCAI*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*.

Morgado, P., & Vasconcelos, N. (2017). Semantically consistent regularization for zero-shot recognition. *In: CVPR*.

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S., & Dean, J. (2014). Zero-shot learning by convex combination of semantic embeddings. *ICLR*.

Patterson, G., Xu, C., Su, H., & Hays, J. (2014). The sun attribute database: Beyond categories for deeper scene understanding. *IJCV*.

Pennington, J., Socher, R., & Manning, C.D. (2014). Glove: Global vectors for word representation. *In: EMNLP*.

Radovanović, M., Nanopoulos, A., & Ivanović, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data.*JMLR*.

Reed, S., Akata, Z., Lee, H., & Schiele, B. (2016). Learning deep representations of fine-grained visual descriptions. *In: CVPR*.

Romera-Paredes, B., & Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. *In: ICML*.

Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., & Matsumoto, Y. (2015). Ridge regression, hubness, and zero-shot learning.*In: ECML*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*.

Song, J., Shen, C., Yang, Y., Liu, Y., & Song, M. (2018). Transductive unbiased embedding for zero-shot learning.*In: CVPR*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions.*In: CVPR*.

Vyas, M.R., Venkateswara, H., & Panchanathan, S. (2020). Leveraging seen and unseen semantic relationships for generative zero-shot learning. arXiv preprint arXiv:2007.09549.

Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

Wan, Z., Chen, D., Li, Y., Yan, X., Zhang, J., Yu, Y., et al. (2019). Transductive zero-shot learning with visual structure constraint. *Advances in Neural Information Processing Systems*, *32*, 9972–9982.

Wang, W., Pu, Y., Verma, V.K., Fan, K., Zhang, Y., Chen, C., Rai, P., & Carin, L. (2018). Zero-shot learning via class-conditioned deep generative models. *AAAI*.

Wang, X., Ye, Y., & Gupta, A. (2018). Zero-shot recognition via semantic embeddings and knowledge graphs. *In: CVPR*.

Xian, Y., Lampert, C.H., Schiele, B., & Akata, Z. (2018). Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*.

Xian, Y., Lorenz, T., Schiele, B., & Akata, Z. (2018). Feature generating networks for zero-shot learning. *In: CVPR*.

Xian, Y., Sharma, S., Schiele, B., & Akata, Z. (2019). f-vaegan-d2: A feature generating framework for any-shot learning. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10275–10284*.

Ye, M., & Guo, Y. (2017). Zero-shot classification with discriminative semantic representation learning. *In: CVPR*.

Zhang, L., Xiang, T., & Gong, S. (2017). Learning a deep embedding model for zero-shot learning. *In: CVPR*.

Zhang, L., Xiang, T., & Gong, S. (2017). Learning a deep embedding model for zero-shot learning. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2021–2030*.

Zhang, Z., & Saligrama, V. (2015). Zero-shot learning via semantic similarity embedding. *In: ICCV*.

Zhang, Z., & Saligrama, V. (2016). Zero-shot learning via joint latent similarity embedding. *In: CVPR*.

Zhang, Z., & Saligrama, V. (2016). Zero-shot recognition via structured prediction. *In: ECCV*.

Zhao, A., Ding, M., Guan, J., Lu, Z., Xiang, T., & Wen, J.R. (2018). Domain-invariant projection learning for zero-shot recognition. *In: Advances in Neural Information Processing Systems, pp. 1019–1030*.

Zhu, X., & Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation. *Technical Report*, Carnegie Mellon University.

Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., & Elgammal, A. (2018). A generative adversarial approach for zero-shot learning from noisy texts. *In: CVPR*.

Zhu, Y., Xie, J., Tang, Z., Peng, X., & Elgammal, A. (2019). Learning where to look: Semantic-guided multi-attention localization for zero-shot learning. arXiv preprint arXiv:1903.00502.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.