Towards Fine-grained Open Zero-shot Learning: Inferring Unseen Visual Features from Attributes

Yang Long
The University of Sheffield
ylong2@sheffield.ac.uk

Li Liu
The University of East Anglia
liuli1213@gmail.com

Ling Shao
The University of East Anglia
ling.shao@ieee.org

Abstract

Zero-shot Learning (ZSL) can leverage attributes to recognise unseen instances. However, the training data is limited and cannot adequately discriminate fine-grained classes with similar attributes. In this paper, we propose a complementary procedure that inversely makes use of attributes to infer discriminative visual features for unseen classes. In this way, ZSL is fully converted into conventional supervised classification, where robust classifiers can be employed to address the fine-grained problem. To infer high-quality unseen data, we propose a novel algorithm named Orthogonal Semantic-Visual Embedding (OSVE) that can discover the tiny visual differences between different instances under the same attribute by an orthogonal embedding space. On two fine-grained benchmarks, CUB and SUN, our method remarkably improves the state-of-the-art results under standard ZSL settings. We further challenge the Open ZSL problem where the number of seen classes is significantly smaller than that of unseen classes. Substantial experiments manifest that the inferred visual features can be successfully fed to SVM which can effectively discriminate unseen classes from fine-grained open candidates.

1. Introduction

Conventional object recognition approaches, such as deep neural networks [12], rely on capturing a large number of training examples paired with annotations to train reliable models. Such a premise can become unattainable in many real-world applications for following reasons: 1) Annotating visual data is expensive. Although numerous images and videos can be freely retrieved from the Internet, the labels of these data are noisy and not competent for existing supervised learning. 2) In the big-data era, the number of new concepts is exponentially increasing, which makes the categories go wider and deeper. It is impractical to collect sufficient visual data for each of the new classes. 3) Capturing instances for rare classes can be infeasible. For example, it is difficult to acquire real photos of ancient or rare species to train a recognition system since the available knowledge could be only textual descriptions or imagined appearances.

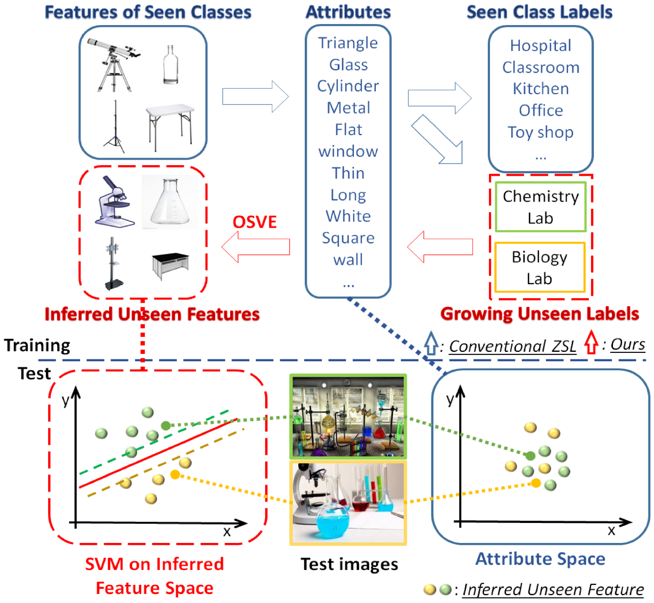

Figure 1. Comparison between our procedure (Red) and the conventional ZSL framework (Blue). Fine-grained classes are often compact and non-describable in the attribute space. Our OSVE can discover tiny visual differences between different instances under the same attribute so as to infer discriminative visual features for unseen classes from fine-grained open candidates.

As a feasible solution, Zero-Shot Learning (ZSL) [15, 13, 23, 33] aims to train semantic models that can generalise to new classes without acquiring unseen visual data at training stage. The standard paradigm of ZSL framework is shown in Fig. 1 (blue path), where a closed-set of seen instances are used to learn a visual-semantic mapping. During the test, images from unseen classes can be firstly mapped to the semantic space and predictions can be made by choosing one of the candidates that are pre-defined by attribute descriptions. However, while new semantics and unseen classes can be incrementally added to the system, the training data is restricted to the closed-set of seen classes without expansion. Under such a framework, there are mainly two problems impeding existing ZSL methods from scaling-up. The first is the *Fine-grained* problem. Namely, the classes are close in the taxonomy, which results in very similar semantic descriptions. Due to existing methods rely on visual-semantic mapping, unseen classes with similar attributes cannot be adequately discriminated. The second is known as the *Open Zero-shot Learning* problem which removes two main unrealistic restrictions of conventional ZSL: 1) all of the candidates for test image must come from unseen classes; 2) the number of seen classes is larger than unseen classes. The first restriction is caused by the correlation problem during attribute designing the results in two attributes A and B may be present or absent together all the time during training. As a result, the test image with only attribute A will be predicted as A& B that is biased towards the seen classes. The second restriction considers the limited size of the training set. Without various seen instances, the learnt semantic model can hardly adapt to unseen classes from a large number of candidates.

In this paper, we propose a complementary approach that inversely infers visual data to train discriminative models for unseen classes. Our method is inspired by the fact that we human can roughly imagine the appearance of unseen objects by associating previous seen classes. Accordingly, as shown in Fig. 1 (Red), our method can inversely infer discriminative visual features from attribute descriptions of unseen instances. In this way, inferred features can be used to train classifiers for unseen classes as conventional supervised learning, *e.g.* SVM. Such a new framework has two potential advantages. Firstly, the training set can be expanded to new unseen classes, which can benefit the open ZSL problem if the number of unseen classes becomes large. Secondly, our classifier is now trained on the original visual feature space without quantisation to the attribute space that is often too compact for the fine-grained problem. For example, the *Biology Lab* and *Chemical Lab* are not discriminative in the semantic space since they share most of the attributes. But, in the visual space, we can enlarge tiny differences between various instances with the same attribute, which, consequently, make fine-grained classes more discriminative.

In spite of that our idea is simple and intuitive, there are two main unsolved technical issues. 1) *Semantic-visual discrepancy*: since attributes are compact high-level representations whereas visual data is usually long-tailed low-level features, the data structure in the two spaces are distinctive. Two close points in the attribute space can be far away in the visual feature space, and vice versa. Due to the structural difference, normal embedding processes are prone to learn the principal components between the two spaces, by which the learnt feature distribution is concentrated and not discriminative. 2) Semantic correlation: like that in the conventional ZSL framework, different attributes may be assigned to the same pattern of visual features. As a result, the inferred unseen features are prone to fall into the clusters of seen features. Considering the above two problems, we propose a novel Orthogonal Semantic-Visual Embedding (OSVE) algorithm to infer visual features from attributes. The key idea is to find an intermediate embedding space that can compromise the structural difference between the visual and semantic space. Meanwhile, we hope to remove the correlations between different attributes, and between seen and unseen classes. To this end, our algorithm jointly optimise the semantic-visual reconstruction error and the orthogonalisation, where the redundancy can be removed in the orthogonal embedding space so that the remaining bases are then decorrelated. We summarise our contribution as follows.

i. We propose to inversely infer discriminative visual features from the attributes unseen classes. Such a framework can make the training set grow with newly added unseen classes in the open ZSL problem. Typical power classifiers, such as SVM, can be employed directly in the feature space rather than the attribute space to improve the fine-grained recognition performance.

ii. We propose a novel OSVE algorithm that can effectively infer visual features and meanwhile remove the correlations. On two benchmarks, our OSVE outperforms state-of-the-art methods under conventional ZSL scenarios

iii. We further challenge two sets of Fine-grained Open ZSL tasks. On both sets of tasks, our OSVE demonstrates promising recognition performance. Extensive experiments manifest that our algorithm can successfully capture the significant visual features from the attributes of unseen classes.

The rest of the paper is organised as follows. In Section 2, we review related ZSL approaches. In Section 3, our algorithm is formalised and introduced. We provide extensive experiments on both conventional and fine-grained open ZSL settings in Section 4. In the last Section 5, we conclude our work and state some possible future work.

2. Related work

We compare our paradigm and that of conventional ZSL in Fig. 1. Most of previous ZSL work is based on (or similar to) the framework called Direct Attribute Prediction (DAP) [13, 14, 20, 38]. For each attribute, a binary classifier is trained using all of the seen classes. During the test, a prediction can be made by Maximum-a-Posteriori criteria over all of the outputs of the binary classifiers. The main drawback of such framework is the correlation problem that reported in [10]. Besides, the human-defined attribute list can be unrealistic and noisy and need to be selected [9, 7, 16, 18]. Therefore, many previous work seeks for an effective form of semantic representation such as class taxonomies [22, 28, 19], or textual features [29, 21]. However, due to other semantic sources cannot provide direct and compact descriptions to the visual appearances, semantic attributes remain the most popular side information for ZSL learning.

A recent trend of ZSL methods adopts the framework of Attribute-Label Embedding (ALE) that jointly estimate all of the attributes by an embedding function from visual to attribute space. Such a framework skilfully avoid the correlation problem or attribute selection since the embedding can optimise the weight of each attribute. Moreover, such a framework be straightforwardly combined with Deep Neraul Network [26]. The much recent research adopts the embedding approach and demonstrates state-of-the-art performance [30, 2, 39, 40, 11, 25]. The remaining challenges so far is to break the restrictions of conventional ZSL settings. [8, 27] focus on transitive settings which view ZSL as a domain adaptation problem. These methods are based on the assumption that unlabelled data of unseen classes can be obtained. Reed *et al.* [26] addresses fine-grained ZSL by a Deep Symmetric Structured Joint Embedding (DA-SJE). Zhang and Saligrama [39] investigate how their method can withstand the reduction of the training set size.

Aside of ALE, some work also considers the drawbacks of direct mapping from visual to semantic spaces. Accordingly, latent attributes [35, 3, 32, 37] aims to discover the statistical relationships between visual and semantic features so as to eliminate the human bias in the attributes. Yu *et al.* [36] use one-to-one classifiers to estimate the similarity of between pair of classes. [3, 17, 31] aim to remove the visual-semantic ambiguity through an intermediate embedding space. [35, 4] proposes bilinear joint embeddings to mitigate the distribution difference between visual and semantic spaces. In [5], classifiers of unseen classes are directly estimated by aligning the manifolds of seen classes.

In comparison to previous methods, our work aims to simultaneously address both fine-grained and open ZSL problems using a unified framework. Our work also adopts attributes as the side information and shares the idea of latent embedding, but our method is inverse and complementary to existing work. While most of the previous methods focus on visual to semantic embedding, our approach focuses on semantic-visual embedding, which is more challenging and requires more consideration. We also consider the imperfection of human-designed attributes, for which we propose a novel orthogonalised embedding approach. The most related work is [6] that attempts to predict visual exemplars for unseen classes. However, their output is a single point in the semantic embedding space, whereas our method can infer instance-level visual features, the number of which equals to that of unseen instances. In short, our unique contribution is to convert ZSL problem into the conventional supervised classification for fine-grained open ZSL using orthogonalised latent embedding.

3. Visual Feature Inference

3.1. Problem setup

The training set contains samples, attributes, and class labels that are in 3-tuples: $(x_1, a_1, y_1), ..., (x_N, a_N, y_N) \subseteq \mathcal{X}_s \times \mathcal{A}_s \times \mathcal{Y}_s$, where N is the number of training samples; $\mathcal{X}_s = [x_{dn}] \in \mathbb{R}^{D \times N}$ is a D-dimensional feature space; $\mathcal{A}_s = [a_{mn}] \in \mathbb{R}^{M \times N}$ is a M-dimensional attribute space; and $y_n \in \{1, ..., C\}$ consists of C discrete class labels. In order to deal with fine-grained open ZSL, we use instance-level attributes, *i.e.* each image is paired with a unique attribute signature. Suppose there are $\hat{N}$ pairs of 'unseen' attributes from $\hat{C}$ discrete classes: $(\hat{a}_1, \hat{y}_1), ..., (\hat{a}_{\hat{N}}, \hat{y}_{\hat{N}}) \subseteq \mathcal{A}_u \times \mathcal{Y}_u$, where $\mathcal{Y}_u \cap \mathcal{Y}_s = \varnothing$, $\mathcal{A}_u = [a_{m\hat{n}}] \in \mathbb{R}^{M \times \hat{N}}$. The goal of zero-shot learning is to learn a classifier, $f : \mathcal{X}_u \to \mathcal{Y}_u$, where the samples in $\mathcal{X}_u$ are completely unavailable during training. We use *Calligraphic* typeface to indicate a space. Subscript s and u refer to 'seen' and 'unseen'. *hat* denotes the variables that are related to 'unseen' samples.

Semantic-Visual Embedding: We aim to infer the visual features of unseen classes by given the semantic attributes. Specifically, we learn a embedding function on the training set $f : \mathcal{A}_s \to \mathcal{X}_s$. After that, we are able to infer $\mathcal{X}_u$ though: $\mathcal{X}_u = f(\mathcal{A}_u)$.

Zero-shot Recognition: Using the inferred visual features, we can directly estimate the probability distribution of the unseen classes. It is straightforward to employ existing supervised classification methods, *i.e.* $f : \mathcal{X}_u \to \mathcal{Y}_u$.

3.2. Orthogonal Semantic-Visual Embedding

Conventional ZSL methods minimise the single classification error of each attribute. Due to the attributes are separately learnt, as aforementioned, such a framework highly depends on the quality of the designed attributes. Recently, there is a new scheme that addresses ZSL by an embedding approach [1]. In particular, an objective function is learnt to simultaneously minimise the multi-class error and also consider the relationship between different attributes. A typical multi-attributes classifier can be formalised as the following problem:

$$\min_{W} \mathcal{L}(W\mathcal{X}_s, \mathcal{A}_s) + \lambda\Omega(W), \tag{1}$$

where W is the mapping matrix, $\mathcal{L}$ is a loss function, and Ω is a regularisation term with its hyper-parameter λ. During the test, an unseen instance can be directly mapped to the attribute space by: $\hat{a} = W\hat{x}$.

However, due to W is learnt using only the training data, the inferred attributes $\hat{a}$ are prone to be biased towards the 'seen' attributes $\mathcal{A}_s$. Since the number of dimension of the visual feature is dominantly large, i.e., $D \gg M$, the mapped semantic data is too compact to distinguish fine-grained unseen classes. Inspired by the idea that a human can imagine the visual appearance of an unseen object through given semantic descriptions, we proposed to infer the visual feature of the unseen classes by reversely learning a mapping function from semantic space to the visual feature space:

$$\min_{W} \mathcal{L}(W\mathcal{A}_s, \mathcal{X}_s) + \lambda\Omega(W). \tag{2}$$

The loss term accounts the reconstruction error between the semantic input and visual output; whereas the regularisation ensures the discrimination to unseen classes. Such a framework provides a direct mapping to the visual space without computing a pseudo-inverse matrix that can lead to information loss. Before the test, it is straightforward to infer the visual features of unseen classes using their class attributes:

$$\mathcal{X}_u = W\mathcal{A}_u. \tag{3}$$

In spite of the simplicity of the above framework, several problems are worth noting. Firstly, in practice, there is often a huge gap between visual and semantic spaces. Compared to the compact attribute representation, the variance of visual data is usually larger due to outliers and noise. Also, the data distribution of the two spaces is distinctive. Thus, directly mapping from semantic to visual space can lead to inferior performance. We propose to insert a latent embedding space $\mathcal{V}$ to reconcile the semantic space with the visual feature space, where $\mathcal{V} = [v_{kn}] \in \mathbb{R}^{K\times N}$, and K is an adjustable number of dimension of $\mathcal{V}$. Secondly, in order to learn discriminative features, we need to remove the correlation between each attribute ao as to ensure better generality. For this purpose, the embedding space should be strictly orthogonal. If we consider a multi-variable linear regression model, the loss function can be defined as:

$$\begin{aligned} J = \quad & \|\mathcal{X}_s - W_1\mathcal{V}\|_F^2 + \|\mathcal{V} - W_2\mathcal{A}_s\|_F^2 \\ & +\lambda\|W_1\|_F^2 + \lambda\|W_2\|_F^2, \textit{s.t. } \mathcal{V}\mathcal{V}^T = I, \end{aligned} \tag{4}$$

where $\|.\|_F$ is the Frobenius norm of a matrix, which estimates the Euclidean distance between two matrices. The latent embedding space $\mathcal{V}$ is decomposed from $\mathcal{X}$, and $\mathcal{A}$ is decomposed from $\mathcal{V}$. $W_1 = [w_{1_{dk}}] \in \mathbb{R}^{D\times K}$ and $W_2 = [w_{2_{km}}] \in \mathbb{R}^{K\times M}$ are embedding matrices. The above Eq. 4 helps us to understand our approach. The embedding space can preserve the principal components between the visual and semantic spaces. Meanwhile, the data structure is scattered so that the inferred features can be discriminative and decorrelated to the original attributes. However, because of the fast decay of eigenvalues, the strict orthogonal constraint can impair the reconstruction of the visual features. Therefore, we relax the constraint. The overall loss function is:

$$\begin{aligned} J = \quad & \|\mathcal{X}_s - W_1\mathcal{V}\|_F^2 + \|\mathcal{V} - W_2\mathcal{A}_s\|_F^2 \\ & +\lambda\|W_1\|_F^2 + \lambda\|W_2\|_F^2 + \beta\|\mathcal{V}\mathcal{V}^T - I\|_F^2. \end{aligned} \tag{5}$$

3.3. Optimisation Strategy

Each term of the above Eq. 5 is convex. However, It is non-convex in $W_1, W_2, \mathcal{V}$ all together. To our best knowledge, there is no direct solution to find the global optima. In this paper, we adopt an alternating optimisation strategy to find the local minima for each term separately. Specifically, the whole task is in turn separated into three sub-problems.

1. W_1**-step:** Suppose we compute the partial derivative of the overall loss function J with respect to W_1, then W_2 and $\mathcal{V}$ are fixed as constants. The loss function becomes a standard least squares problem. Let the partial derivative equal to zero; then we have the closed form solution:

$$\begin{aligned} \min_{W_1} \quad & \|\mathcal{X}_s - W_1\mathcal{V}\|_F^2 + \lambda\|W_1\|_F^2 \\ \frac{\partial J}{\partial W_1} \quad &= \quad -2(W_1\mathcal{V} - \mathcal{X}_s)\mathcal{V}^T + 2\lambda W_1 = 0 \\ W_1 \quad &= \quad \mathcal{X}_s\mathcal{V}^T\left(\mathcal{V}\mathcal{V}^T + \lambda I\right)^{-1}. \end{aligned} \tag{6}$$

2. W_2**-step:** Similar to the step 1, we can fix W_1 and $\mathcal{V}$, and compute the partial derivative of J with respect to W_2. The corresponding solution is:

$$\begin{aligned} \min_{W_2} \quad & \|\mathcal{V} - W_2\mathcal{A}_s\|_F^2 + \lambda\|W_2\|_F^2 \\ \frac{\partial J}{\partial W_2} \quad &= \quad -2(W_2\mathcal{A}_s - \mathcal{V})\mathcal{A}_s^T + 2\lambda W_2 = 0 \\ W_2 \quad &= \quad \mathcal{V}\mathcal{A}_s^T\left(\mathcal{A}_s\mathcal{A}_s^T + \lambda I\right)^{-1}. \end{aligned} \tag{7}$$

3. $\mathcal{V}$**-step:** $\mathcal{V}$ should be solved carefully. Since $\mathcal{V}$ is related to all of the three terms, it balances how accurate can we infer the visual feature and how discriminative can the inferred features generalise to unseen data. We propose to solve $\mathcal{V}$ as an independent sub-problem inside the overall optimisation. Fix W_1 and W_2, we can get the partial loss function J_v for $\mathcal{V}$. We then set the partial derivative respect to $\mathcal{V}$ to zero:

$$\begin{aligned} \min_{\mathcal{V}} J_v = \quad & \|\mathcal{X}_s - W_1\mathcal{V}\|_F^2 + \|\mathcal{V} - W_2\mathcal{A}_s\|_F^2 \\ & +\beta\|\mathcal{V}\mathcal{V}^T - I\|_F^2 \\ \frac{\partial J_v}{\partial \mathcal{V}} = \quad & 2W_1^T(W_1\mathcal{V} - \mathcal{X}_s) + 2(\mathcal{V} - W_2\mathcal{A}_s) \\ & +2\beta(\mathcal{V}\mathcal{V}^T - I)\mathcal{V}. \end{aligned} \tag{8}$$

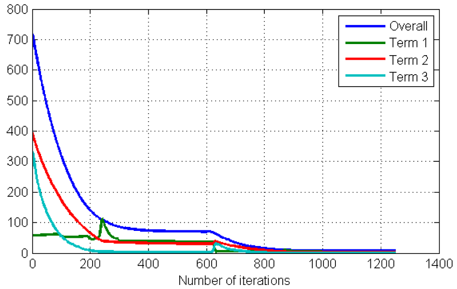

Figure 2. An example of the convergence situations shows the loss with respect to the number of iterations. Term 1 and 2 corresponds to the reconstruction errors to visual and semantic spaces. Term 3 accounts how orthogonal is the embedding space.

Adaptive Gradient Descent: In order to solve the optimal $\mathcal{V}$, we adopt the adaptive gradient decent strategy to solve Eq. 8. We introduce τ to control the learning rate. If J_v keeps converging, τ is increased to accelerate the process. Once J_v becomes diverged, τ is reduced correspondingly to increase the tolerance. Such a strategy is vital for keeping the balance between reconstruction and orthogonalisation. As shown in Fig. 2, the solver firstly focuses on optimising the semantic reconstruction and the orthogonality. After 200 iterations, the learning rate becomes over large that causes the loss of visual reconstruction increased dramatically. Thus, τ is immediately reduced so that the three terms start to be optimised together again. Without such an adaptive scheme, it is unable to control the unpredictable divergence of any of the terms. The whole learning procedure is summarised in Algorithm 1.

$$\mathcal{V}_{t+1} = \mathcal{V}_t - \tau \frac{\partial J_v}{\partial \mathcal{V}} \tag{9}$$

$$\tau_{t+1} = \begin{cases} 1.2\tau \text{ if } J_{v_{t+1}} < J_v \\ 0.5\tau \text{ otherwise} \end{cases}. \tag{10}$$

3.4. Zero-shot Recognition

Once we obtain the embedding matrices W_1 and W_2, the visual features of unseen classes can be easily inferred from their attributes:

$$\mathcal{X}_u = W_1 * W_2 * \mathcal{A}_u. \tag{11}$$

It is noticeable that for instance-level attributes, $\mathcal{X}_u$ contains as many instances as the test set. The zero-shot recognition task now becomes a conventional classification problem. Thus, any existing supervised classifier, *e.g.* SVM, can be applied. Since we focus on the quality of the inferred features in this paper, we compare NN to SVM s well. For NN approach, given a test unseen instance $\hat{x}$, we can predict its class label $\hat{c}$ by:

$$\hat{c} = \arg\min_{c} \|\hat{x} - x_{\hat{n}}\|^2, \text{ where } x_{\hat{n}} \in \mathcal{X}_u, y_{\hat{n}} = c \in \mathcal{Y}_u. \tag{12}$$

Algorithm 1: : OSVE

```
Input: $\{\mathcal{X}, \mathcal{A}, \mathcal{Y}\}, K, \lambda, \beta, \tau$.
Output: $W_1$, and $W_2$.
Initialisation: random initial matrix $\mathcal{V}$.
while Rq. 5 is not converged do
    Update $W_1$ by Eq. 6;
    Update $W_2$ by Eq. 7;
    while Eq. 8 is not converged do
        Update $\mathcal{V}$ by Eq. 9;
        Update $\tau$ by Eq. 10;
    end while
end while
return $W_1$, and $W_2$;
```

Table 1. Key statistics of CUB and SUN datasets.

Dataset	**CUB**	**SUN**
# of Attributes	312	102
Attribute Type	Binary	Continues
Annotation Level	per image & per class	per image
# of Total Images	11788	13430
Seen/Unseen Split	150/50	707/10

4. Experiments

We first introduce the datasets, on which we compare our approach to existing state-of-the-art methods. Since the published results are obtained on different settings, in terms of visual features, seen/unseen splits, and semantic side information, we aim to provide a fair comprehensive comparison to most of the outstanding models. We also provide detailed self-comparisons to baseline methods so as to verify the claims we made in this paper. Finally, we investigate our method on the fine-grained open ZSL tasks.

4.1. Setup

Datasets and Settings Our method is evaluated on two fine-grained datasets, Caltech-UCSD Birds (CUB) [34] and SUN attribute (SUN) [24]. We summarise the key statistics in Table 1. For CUB, there are 11788 images from 200 classes of birds. Many bird species can be hardly differentiated by humans. The usual Seen/Unseen split for ZSL is 150/50. For SUN, the number of classes is 717, which is larger than that of CUB. The total number of images is 13430. Some classes are close on both semantic meanings and visual appearances, *e.g. theatre* and *ballroom*.

Visual Features Existing methods differ in adopted visual features. To make a comprehensive comparison, we implement our method using both shallow features that are released by the datasets and deep features extracted using VGG-19 and released by [39].

Semantic Attributes Both of the datasets now provide instance-level attributes. Each test image is paired with a

Table 2. Comparison to state-of-the-art methods for both datasets. Results are overall accuracies in %.

Caltech-UCSD Birds				**SUN attribute**		
Methods	SI	Shallow features	Deep features	Methods	Shallow features	Deep features
DAP[13]	A	10.5	31.4	DAP[13]	52.50	72.00
AHLE[1]	A+H	18.0	27.3	ZSRwUA[9]	56.18	-
SJE[2]	A+W+H	19.0	47.1	ESEZL[30]	65.75	82.10
UDA[11]	A+W	28.1	40.6	SSE[39]	-	82.50
DS-SJE[26]	A+W+H	-	56.8	JLSE[40]	-	**83.83**
OSVE+NN	A	20.2	45.2	OSVE+NN	56.96	76.21
OSVE+SVM	A	**28.9**	**60.1**	**OSVE+SVM**	**70.59**	83.23

SI: side informations, A: attributes, H: hierarchy, W: word2vec.

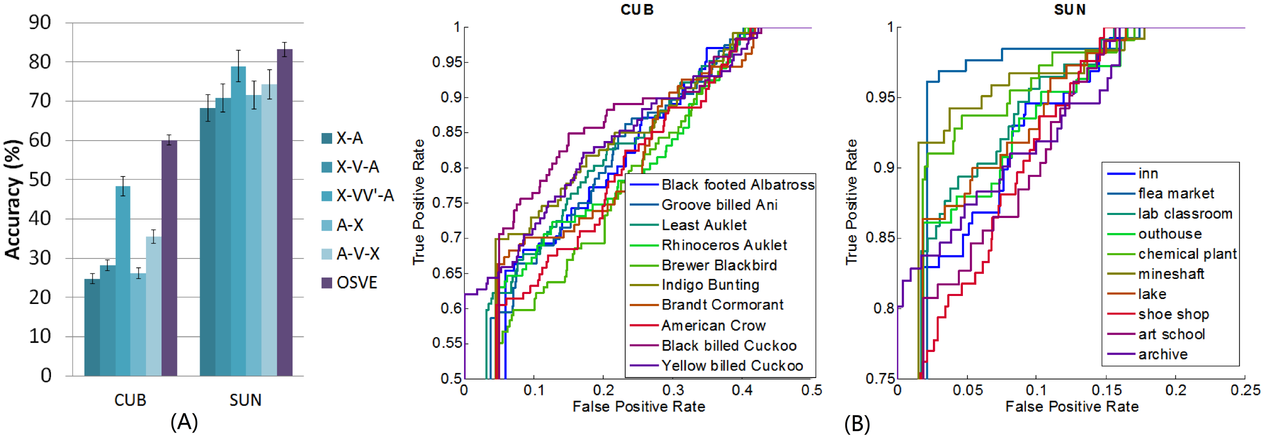

Figure 3. A. overall accuracies of baseline methods by substituting key components of the proposed framework. B. ROC curves of our method on the two datasets. For clarity, only 10 of the 50 unseen classes on CUB are shown.

unique attribute signature based on the actual visual appearance, which is different from the class-level attributes that let all of the images in a class share the same attribute signature. Our method benefits from such a scenario for open ZSL for the reason that, if the number of training classes is small, our algorithm can still discover the differences between instances under the same attribute.

Zero-shot Cross-validation We obtain the optimal hyperparameters through a new cross-validation strategy. Since we aim to address ZSL problems, traditional cross-validation for multi-label classification is not helpful because all of the seen classes are used for both training and validation. Therefore, we propose a novel leave-one-fold-out strategy. The seen classes are divided into ten disjointed folds. We use one fold as unseen validation set and train models on remaining folds. We choose the set of hyperparameters which can lead to the highest mean accuracy on all of the ten folds. We fix this set of parameters for the following experiments.

4.2. Benchmark Comparison

Comparison to State-of-the-art Methods We first compare to previous published results. Due to few methods are evaluated on both of the datasets, separate the results by the two datasets. We summarise our comparison in Table 2.

For CUB, we compare to five methods. DAP [13] is the most common ZSL framework that trains binary SVM classifier for each attribute separately and makes a prediction by Maximum-a-Posteriori. AHLE [1] and SJE [2] both adopt a bilinear compatibility function to make visual to semantic embedding using hierarchical information. But SJE incorporates 1K-dim GoogleNet and textural features. DS-SJE [26] use deep learning to substitute the embedding function and gives state-of-the-art results. UDA [11] views ZSL as a domain adoption problem. Although their setting is slightly different that uses unlabelled unseen data, we still make a comparison because we use inferred unseen data for classification. For both shallow and deep features, our method achieves significant improvements over all of the published results. It is noticeable that our method only uses attributes as side information. Also, the results of using Nearest Neighbour (NN) classifier are slightly lower than that of using SVM, which is caused by that the inferred features become more discriminative after the orthogonal embedding. However, the data structure can be slightly different to real distribution.

For SUN, DAP is also compared. ZSLwUA [9] considers the unreliability of human-defined attributes and make predictions by random forest. ESEZL [30] combines visual-attribute and attribute-label embedding into one joint func-

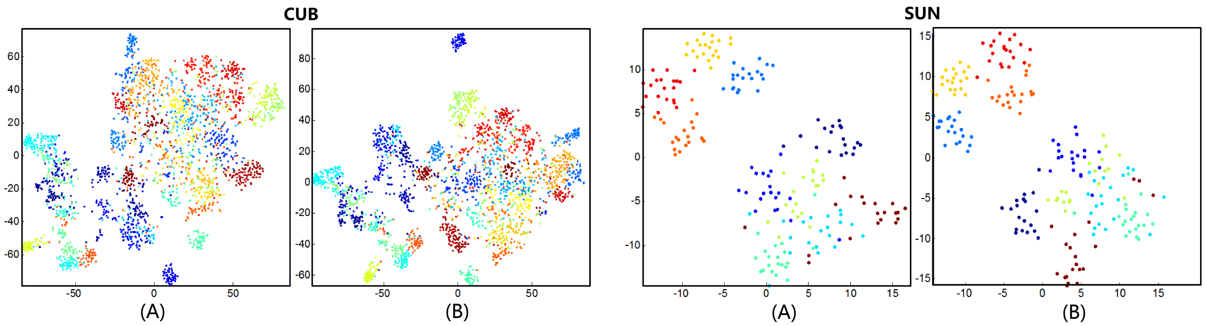

Figure 4. Comparing the data distribution between real (A) and inferred (B) visual features of unseen classes. Note that t-SNE can result in slight distortion and colour differences.

tion. SSE [39] and JLSE [40] are similarity-based approaches that jointly learn a dictionary learning function for both visual and attribute domains. Note that all of the compared methods use attributes as side information. Using deep features, ESEZL, SSE, and JLSE achieves state-of-the-art results. Our result is only 0.5% lower than that of JLSE. However, using shallow features, our method is 5% higher than other methods. Again, we observe that using SVM can significantly boost the performance, which benefits from using inferred visual features.

Fig. 3 (B) depicts the resulting ROC curves of our results on the two datasets. One can see that the performances on all classes are balanced and reasonable.

Analysis To understand how each part of our approach contributes to the overall performance, we also implement a set of baseline methods. We summarise the results in Fig. 3 (A). All of the baseline methods are implemented using deep features. The first three baselines examine the conventional visual-attribute embeddings. We train SVM using the attributes of unseen instances. During the test, images are mapped to the attribute space and classified by the trained attribute-SVM. X-A directly learns a mapping from visual features. X-VV'-A is the inverse version of the proposed method, where we insert an intermediate latent embedding spaces with orthogonal constraints. To see the effect of orthogonality, we remove the orthogonal constraint in X-V-A as a reference. Similarly, for the later three methods using attribute-visual embedding, we compare to A-X that directly maps attributes to the visual space without orthogonalised embedding space. A-V-X is implemented by removing the orthogonal constraint in Equation 5.

We observe the orthogonality contributes the most to the overall performance. Also, embedding from attribute to visual space significantly boosts the performance, which verified our statements that, fine-grained classes are more discriminable in the visual space due to the semantic representations are too close. Another conclusion can be made that inserting an intermediate embedding space is helpful to compromise the data structural differences to some extents. Although without orthogonal constraints, the results of X-V-A and A-V-X are higher than that without the $\mathcal{V}$ space.

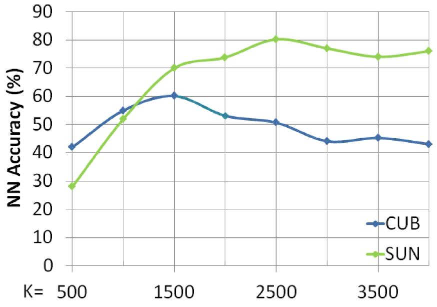

Figure 5. the performance curve respects to the dimension K of the intermediate embedding space.

How many dimensions do we need for $\mathcal{V}$? Since orthogonalisation can effectively remove the redundant information, each dimension of the orthogonal space indicate a reliable component. In Fig. 5, we show the recognition rates vary with respect to the dimension K of the embedding space for the two datasets. It can be seen that best results are given with K equals to 1500 and 2500 respectively. Since the classes in SUN are more various than that in CUB, higher dimensional $\mathcal{V}$ can give better results in general.

Data Distribution of inferred Visual Features One of the fundamental questions is whether our inferred visual features are close to the real data. In Fig. 4, we demonstrate the data distribution of real and inferred visual features using t-SNE. Although t-SNE can result in slight distortion and colour changes, we can still recognise the data structures are preserved. The only difference is that some of the inferred visual features are shown further than the real data. For example, the blue cluster of points at the top in CUB is pulled further by t-SNE, which is because our OSVE can reduce the correlations and make the inferred data more discriminative.

4.3. Fine-grained Open Zero-shot Learning

There are two restrictions for conventional ZSL settings that are not realistic. 1) The test images can only come from unseen class. 2) The number of seen class is substantially

Table 3. Results (in %) of Open ZSL 1: add extra seen classes as candidates or add instances from seen classes for testing.

Dataset	#Extra Seen	For Candidate	Add to Test
CUB	50	56.5	51.9
	100	52.7	43.2
	150	47.1	36.8
	0	60.1	
SUN	10	79.98	76.63
	100	74.38	70.47
	300	65.53	59.81
	500	61.72	54.26
	707	58.42	49.59
	0	83.23	

larger than that of unseen classes. By breaking the restrictions, we investigate two scenarios of open zero-shot learning, both of which widely exist in real-world applications. **Scenario 1:** *Test images come from a mixture of seen and unseen classes.* **Scenario 2:** *Testing by a large number of unseen classes using a small training set.*

For scenario 1, the seen/unseen splits are the same (150/50 for CUB and 707/10 for SUN). But we use half of each seen class for training and the other half for testing. Before the test, we infer the visual features for both seen and unseen test images, using which we train SVM classifiers. In this way, the seen classes are added as candidates, *i.e.* test unseen image now may be misclassified to seen classes. We also add images from seen classes for testing. The potential challenge is that the seen classes may be misclassified into unseen classes. We summarise our results in Table 3. We show the results of conventional ZSL (0 extra seen) as references. It can be seen that by testing on the whole datasets (200 classes in CUB and 717 classes in SUN), our method can still lead to acceptable results.

For scenario 2, we investigate how our method can withstand a significant reduction of seen class number and an increasing unseen class number. Our results are summarised in Fig. 6. Results using a various size of training sets are shown in different colours of lines. We gradually add remaining classes as unseen classes for testing and see the trend of overall recognition rates. We observe the result on the most extreme splits (10/190) on CUB is only 8% lower than that of 10/50. For SUN, increasing the number of unseen classes from 10 to 100 only result in 15% recognition drop in average. Under the extreme setting on SUN (10/707), we achieve 22.4% recognition rate, where the random guess is only 0.14%.

Qualitative Results As shown in Fig. 7, given a query unseen instance, we infer its visual feature and examine what do the original images of the nearest features look like. We compare the results under conventional and extreme open ZSL settings. It can be seen that the tasks are difficult even for humans. The inferred visual features can still retrieve the most visually similar instances.

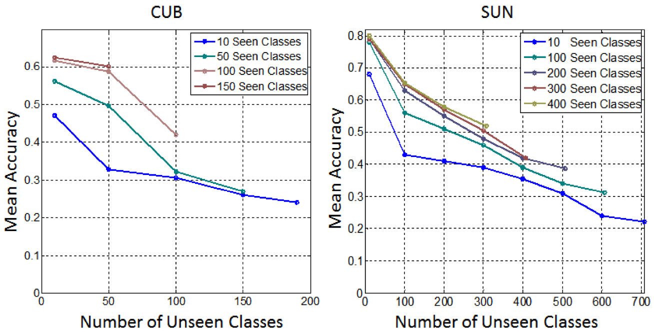

Figure 6. Open ZSL 2: test by increasing number of unseen classes using different size of training sets.

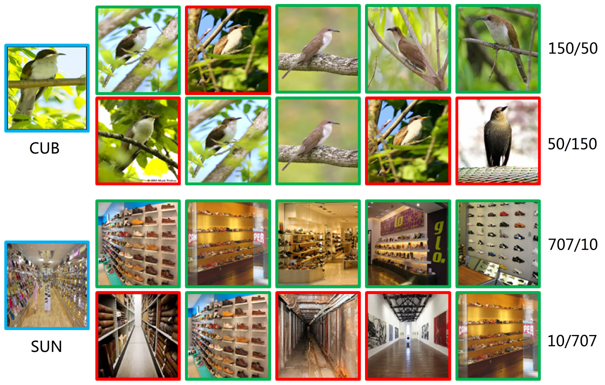

Figure 7. Top-5 nearest neighbours of the query image under conventional and open ZSL. Correct and incorrect matches are shown in green and red respectively. Corresponding seen/unseen splits are shown on the right.

5. Conclusion

In this paper, we proposed a novel semantic-visual embedding framework that was inverse to conventional ZSL frameworks . Using inferred visual features, we could convert the ZSL problem into conventional supervised classification and employ powerful classifiers for fine-grained open ZSL. On standard seen/unseen settings, our method achieved significant improvements over the state-of-the-art results. Furthermore, we challenged two scenarios of open ZSL tasks, on both of which our method manifested promising performance. Also, the inferred visual features were shown under the same data distribution as real data. We ascribe the success of our method to the orthogonal embedding space that can jointly compromise the structural differences between visual and attribute spaces and remove the redundant correlations simultaneously.

For future work, our method is helpful to synthesise visual data for rare unseen classes. Our method can also be applied to incremental ZSL frameworks that can mutually infer new attributes and visual data in a large-scale recognition system.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015.

[3] Z. Al-Halah, T. Gehrig, and R. Stiefelhagen. Learning semantic attributes via a common latent space. In *VISAPP*, 2014.

[4] Z. Cai, L. Liu, M. Yu, and L. Shao. Latent structure preserving hashing. In *BMVC*, 2015.

[5] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016.

[6] S. Changpinyo, W.-L. Chao, and F. Sha. Predicting visual exemplars of unseen classes for zero-shot learning. *arXiv preprint arXiv:1605.08151*, 2016.

[7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[8] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Learning multimodal latent attributes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(2):303–316, 2014.

[9] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *NIPS*, 2014.

[10] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, 2014.

[11] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[13] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[14] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(3):453–465, 2014.

[15] H. Larochelle, D. Erhan, and Y. Bengio. Zero-data learning of new tasks. In *AAAI*, 2008.

[16] Y. Long, L. Liu, and L. Shao. Attribute embedding with visual-semantic ambiguity removal for zero-shot learning. In *BMVC*, 2016.

[17] Y. Long, L. Liu, and L. Shao. Attribute embedding with visual-semantic ambiguity removal for zero-shot learning. In *BMVC*, 2016.

[18] Y. Long and L. Shao. Describing unseen classes by exemplars: Zero-shot learning using grouped simile ensemble. In *WACV*, 2017.

[19] Y. Long, F. Zhu, and L. Shao. Recognising occluded multi-view actions using local nearest neighbour embedding. *Computer Vision and Image Understanding*, 144:36–45, 2016.

[20] D. Mahajan, S. Sellamanickam, and V. Nair. A joint learning framework for attribute models and object descriptions. In *ICCV*, 2011.

[21] T. Mensink, E. Gavves, and C. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014.

[22] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *ECCV*. 2012.

[23] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

[24] G. Patterson, C. Xu, H. Su, and J. Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014.

[25] J. Qin, Y. Wang, L. Liu, J. Chen, and L. Shao. Beyond semantic attributes: Discrete latent attributes learning for zero-shot recognition. *IEEE Signal Processing Letters*, 23(11):1667–1671, 2016.

[26] S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016.

[27] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *NIPS*, 2013.

[28] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2011.

[29] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010.

[30] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015.

[31] L. Shao, L. Liu, and M. Yu. Kernelized multiview projection for robust action recognition. *International Journal of Computer Vision*, 118(2):115–129, 2016.

[32] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Augmented attribute representations. In *ECCV*. 2012.

[33] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[34] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[35] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016.

[36] F. Yu, L. Cao, R. Feris, J. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013.

[37] M. Yu, L. Liu, and L. Shao. Structure-preserving binary representations for rgb-d action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 38(8):1651–1664, 2016.

[38] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *ECCV*. 2010.

[39] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015.

[40] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016.